

# LEVEL SET METHODS FOR FINDING CRITICAL POINTS OF MOUNTAIN PASS TYPE

A.S. LEWIS AND C.H.J. PANG

**ABSTRACT.** Computing mountain passes is a standard way of finding critical points. We describe a numerical method for finding critical points that is convergent in the nonsmooth case and locally superlinearly convergent in the smooth finite dimensional case. We apply these techniques to describe a strategy for the Wilkinson problem of calculating the distance of a matrix to a closest matrix with repeated eigenvalues. Finally, we relate critical points of mountain pass type to nonsmooth and metric critical point theory.

## CONTENTS

1. Introduction	1
2. A level set algorithm	4
3. A locally superlinearly convergent algorithm	6
4. Superlinear convergence of the local algorithm	7
5. Further properties of the local algorithm	16
6. Saddle points and criticality properties	22
7. Wilkinson's problem: Background	26
8. Wilkinson's problem: Implementation and numerical results	28
9. Non-Lipschitz convergence and optimality conditions	29
Acknowledgments	34
References	34

**Keywords:** mountain pass, nonsmooth critical points, superlinear convergence, metric critical point theory, Wilkinson distance.

## 1. INTRODUCTION

Computing mountain passes is an important problem in computational chemistry and in the study of nonlinear partial differential equations. We begin with the following definition.

**Definition 1.1.** Let  $X$  be a topological space, and consider  $a, b \in X$ . For a function  $f : X \rightarrow \mathbb{R}$ , define a *mountain pass*  $p^* \in \Gamma(a, b)$  to be a minimizer of the problem

$$\inf_{p \in \Gamma(a, b)} \sup_{0 \leq t \leq 1} f \circ p(t).$$

Here,  $\Gamma(a, b)$  is the set of continuous paths  $p : [0, 1] \rightarrow X$  such that  $p(0) = a$  and  $p(1) = b$ .

---

*Date:* June 22, 2010.

An important problem in computational chemistry is to find the lowest energy to transition between two stable states. If  $a$  and  $b$  represent two states and  $f$  maps the states to their potential energies, then the mountain pass problem calculates this lowest energy. Early work on computing transition states includes Sinclair and Fletcher [38], and recent work is reviewed by Henkelman, Jóhannesson and Jónsson [21]. We refer to this paper for further references in the Computational Chemistry literature.

Perhaps more importantly, the mountain pass idea is also a useful tool in the analysis of nonlinear partial differential equations. For a Banach space  $X$ , variational problems are problems (P) such that there exists a smooth functional  $J : X \rightarrow \mathbb{R}$  whose critical points (points where  $\nabla J = 0$ ) are solutions of (P). Many partial differential equations are variational problems, and critical points of  $J$  are “weak” solutions. In the landmark paper by Ambrosetti and Rabinowitz [4], the mountain pass theorem gives a sufficient condition for the existence of critical points in infinite dimensional spaces. If an optimal path to solve the mountain pass problem exists and the maximum along the path is greater than  $\max(f(a), f(b))$ , then the maximizer on the path is a critical point distinct from  $a$  and  $b$ . The mountain pass theorem and its variants are the primary ways to establish the existence of critical points and to find critical points numerically. For more on the mountain pass theorem and some of its generalizations, we refer the reader to [24].

In [13], Choi and McKenna proposed a numerical algorithm for the mountain pass problem by using an idea from Aubin and Ekeland [5] to solve a semilinear partial differential equation. This is extended to find solutions of *Morse index 2* (that is, the maximum dimension of the subspace of  $X$  on which  $J''$  is negative definite) in Ding, Costa and Chen [19], and then to higher Morse index by Li and Zhou [26].

Li and Zhou [27], and Yao and Zhou [45] proved convergence results to show that their minimax method is sound for obtaining weak solutions to nonlinear partial differential equations. Moré and Munson [33] proposed an “elastic string method”, and proved that the sequence of paths created by the elastic string method contains a limit point that is a critical point.

The prevailing methods for numerically solving the mountain pass problem are motivated by finding a sequence of paths (by discretization or otherwise) such that the maximum along these paths decrease to the optimal value. Indeed, many methods in [21] approximate a mountain pass in this manner. As far as we are aware, only [6, 22] deviate from this strategy. We make use of a different approach by looking at the path connected components of the lower level sets of  $f$  instead.

One easily sees that  $l$  is a lower bound of the mountain pass problem if and only if  $a$  and  $b$  lie in two different path connected components of  $\text{lev}_{\leq l} f$ . A strategy to find an optimal mountain pass is to start with a lower bound  $l$  and keep increasing  $l$  until the path connected components of  $\text{lev}_{\leq l} f$  containing  $a$  and  $b$  respectively coalesce at some point. However, this strategy requires one to determine whether the points  $a$  and  $b$  lie in the same path connected component, which is not easy. We turn to finding saddle points of mountain pass type, as defined below.

**Definition 1.2.** For a function  $f : X \rightarrow \mathbb{R}$ , a *saddle point of mountain pass type*  $\bar{x} \in X$  is a point such that there exists an open set  $U$  such that  $\bar{x}$  lies in the closure of two path components of  $(\text{lev}_{< f(\bar{x})} f) \cap U$ .

We shall refer to saddle points of mountain pass type simply as saddle points. As an example, for the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  defined by  $f(x) = x_1^2 - x_2^2$ , the point  $\mathbf{0}$  is a saddle point of mountain pass type: We can choose  $U = \mathbb{R}^2$ ,  $a = (0, 1)$ ,  $b = (0, -1)$ . When  $f$  is  $\mathcal{C}^1$ , it is clear that saddle points are critical points. As we shall see later (in Propositions 6.1 and 6.2), saddle points of mountain pass type can, under reasonable conditions, be characterized as maximal points on mountain passes, acting as “bottlenecks” between two components. In fact, if  $f$  is  $\mathcal{C}^2$ , the Hessians are nonsingular and several mild assumptions hold, these bottlenecks are exactly critical points of Morse index 1. We refer the reader to the lecture notes by Ambrosetti [3]. Some of the methods in [21] actually find saddle points instead of solving the mountain pass problem.

We propose numerical methods to find saddle points using the strategy suggested in Definition 1.2. We start with a lower bound  $l$  and keep increasing  $l$  until the components of the level set  $\text{lev}_{\leq l} f \cap U$  containing  $a$  and  $b$  respectively coalesce, reaching the objective of the mountain pass problem. The first method we propose in Algorithm 2.1 is purely metric in nature. One appealing property of this method is that calculations are now localized near the critical point and we keep track of only two points instead of an entire path. Our algorithm enjoys a monotonicity property: The distance between two components decreases monotonically as the algorithm progresses, giving an indication of how close we are to the saddle point. In a practical implementation, local optimality properties in terms of the gradients (or generalized gradients) can be helpful for finding saddle points. Such optimality conditions are covered in Section 9.

It follows from the definitions that our algorithm, if it converges, converges to a saddle point. We then prove that any saddle point is deformationally critical in the sense of metric critical point theory [17, 25, 23], and is Morse critical under additional conditions. This implies in particular that any saddle point is Clarke critical in the sense of nonsmooth critical point theory [12, 37] based on nonsmooth analysis in the spirit of [8, 14, 32, 36]. It seems that there are few existing numerical methods for finding either critical points in a metric space or nonsmooth critical points. Currently, we are only aware of [44].

One of the main contributions of this paper is to give a second method (in Section 3) which converges locally superlinearly to a nondegenerate smooth critical point, i.e., critical points where the Hessian is nonsingular, in  $\mathbb{R}^n$ . A potentially difficult step in this second method is that we have to find the closest point between two components of the level sets. While the effort needed to perform this step accurately may be great, the purpose of this step is to make sure that the problem is well aligned after this step. Moreover, this step need not be performed to optimality. In our numerical example in Section 8, we were able to obtain favorable results without performing this step.

Our initial interest in the mountain pass problem came from computing the 2-norm distance of a matrix  $A$  to the closest matrix with repeated eigenvalues. This is also known as the Wilkinson problem, and this value is the smallest 2-norm perturbation that will make the eigenvalues of matrix  $A$  behave in a non-Lipschitz manner. Alam and Bora [1] showed how the Wilkinson’s problem can be reduced to a global mountain pass problem. We do not solve the global mountain pass problem associated with the Wilkinson problem, but we demonstrate that locally our algorithm converges quickly to a smooth critical point of mountain pass type.

**Outline:** Section 2 illustrates a local algorithm to find saddle points of mountain pass type, while Sections 3, 4 and 5 are devoted to the statement, proof of convergence, and additional observations of a fast local algorithm to find nondegenerate critical points of Morse index 1 in  $\mathbb{R}^n$ .

Sections 6 discusses the relationship between mountain passes, saddle points, and critical points in the sense of metric critical point theory and nonsmooth analysis, and does not depend on material in Sections 3, 4 and 5.

Finally, Sections 7 and 8 illustrates the fast local algorithm in Section 3. Section 9 discusses optimality conditions for the subproblem in the algorithm in Section 2.

**Notation:** As we will encounter situations where we want to find the square of the  $j$ th coordinate of the  $i$ th iterate of  $x$ , we write  $x_i^2(j)$  in the proof of Theorem 4.8. In other parts, it will be clear from context whether the  $i$  in  $x_i$  is used as an iteration counter or as a reference to the  $i$ th coordinate. Let  $\mathbb{B}^d(\mathbf{0}, r)$  be the ball with center  $\mathbf{0}$  and radius  $r$  in  $\mathbb{R}^d$ , and  $\mathring{\mathbb{B}}^d(\mathbf{0}, r)$  be the corresponding open ball.

## 2. A LEVEL SET ALGORITHM

We present a level set algorithm to find saddle points. Assume  $f : X \rightarrow \mathbb{R}$ , where  $(X, d)$  is a metric space.

**Algorithm 2.1.** (*Level set algorithm*) *A local bisection method for approximating a mountain pass from  $x_0$  to  $y_0$  for  $f|_U$ , where both  $x_0$  and  $y_0$  lie in some open path connected set  $U$ .*

- (1) Start with an upper bound  $u$  and a lower bound  $l$  for the objective of the mountain pass problem and  $i = 0$ .
- (2) Solve the optimization problem

$$(2.1) \quad \begin{array}{ll} \min & d(x, y) \\ \text{s.t.} & x \in S_1, y \in S_2 \end{array}$$

where  $S_1$  is the component of the level set  $(\text{lev}_{\leq \frac{1}{2}(l+u)} f) \cap U$  that contains  $x_i$  and  $S_2$  is the component that contains  $y_i$ .

- (3) If  $S_1$  and  $S_2$  are the same component, then  $\frac{1}{2}(l+u)$  is an upper bound, otherwise it is a lower bound. Update the upper and lower bounds accordingly. In the case where the lower bound is changed, increase  $i$  by 1, and let  $x_i$  and  $y_i$  be the minimizers of (2.1). For future discussions, let  $l_i$  corresponding value of  $l$  to  $x_i$  and  $y_i$ . Repeat step 2 until  $x_i$  and  $y_i$  are sufficiently close.
- (4) If an actual approximate mountain pass is desired, take a path  $p_i : [0, 1] \rightarrow U \cap (\text{lev}_{\leq u} f)$  connecting the points

$$x_0, x_1, \dots, x_{i-2}, x_{i-1}, x_i, y_i, y_{i-1}, y_{i-2}, \dots, y_1, y_0.$$

Step (3) is illustrated in Figure 2.1.

To start the algorithm, an upper bound  $u$  can be taken to be the maximum of any path from  $x_0$  to  $y_0$ , while a lower bound can be the maximum of  $f(x_0)$  and  $f(y_0)$ . In fact, in step (3), we may update the upper bound  $u$  to be the maximum along the line segment joining  $x_i$  and  $y_i$  if it is a better upper bound.

In practice, one need not solve subproblem (2.1) in step 2 too accurately, as it might be more profitable to move on to step 3. While theory demands the global optimizers for subproblem (2.1), an implementation of Algorithm 2.1 can only find

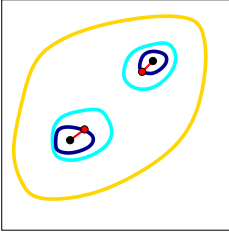
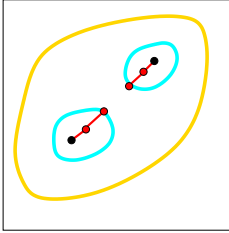
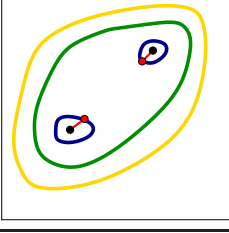
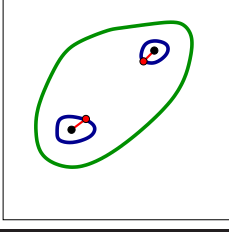
Case	Before	After
$\{x \mid f(x) \leq \frac{u+l}{2}\}$ 2 components		
1 component		

FIGURE 2.1. Illustration of Algorithm 2.1.

local optimizers, which is not sufficient for the global mountain pass problem, but can be successful for the purpose of finding saddle points. The optimality conditions in terms of gradients (or generalized gradients) can be helpful for characterizing local optimality (see Section 9). Notice that the saddle point property is local. If  $x_i$  and  $y_i$  converge to a common limit, then it is clear from the definitions that the common limit is a saddle point.

Another issue with subproblem (2.1) in step 2 is that minimizers may not exist. For example, the sets  $S_1$  and  $S_2$  may not be compact. We now discuss how convergence to a critical point in Algorithm 2.1 can fail in the finite dimensional case.

The Palais-Smale condition is important in nonlinear analysis, and is often a necessary condition in the smooth and nonsmooth mountain pass theorems and other critical point existence theorems. We refer to [29, 34, 35, 39, 42] for more details. We recall its definition.

**Definition 2.2.** Let  $X$  be a Banach space and  $f : X \rightarrow \mathbb{R}$  be a  $\mathcal{C}^1$  functional. We say that a sequence  $\{x_i\}_{i=1}^\infty \subset X$  is a *Palais-Smale sequence* if  $\{f(x_i)\}_{i=1}^\infty$  is bounded and  $f'(x_i) \rightarrow \mathbf{0}$ , and  $f$  satisfies the *Palais-Smale condition* if any Palais-Smale sequence admits a convergent subsequence.

For nonsmooth  $f$ , the condition  $f'(x_i) \rightarrow \mathbf{0}$  is  $\inf_{x_i^* \in \partial f(x_i)} |x_i^*| \rightarrow 0$  instead.

In the absence of the Palais-Smale condition, Algorithm 2.1 may fail to converge because the sequence  $\{(x_i, y_i)\}_{i=1}^\infty$  need not have a limit point of the form  $(\bar{z}, \bar{z})$ , or the sequence  $\{(x_i, y_i)\}_{i=1}^\infty$  need not even exist. The examples below document the possibilities.

**Example 2.3.** (a) Consider  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  defined by  $f(x, y) = e^{-x} - y^2$ . Here, the distance between the two components of the level sets is zero for all  $\text{lev}_{\leq c} f$ , where  $c < 0$ , and  $x_i$  and  $y_i$  do not exist. The sequence  $\{(i, 0)\}_{i=1}^\infty$  is a Palais-Smale sequence but does not converge.

(b) For  $f(x, y) = e^{-2x} - y^2 e^{-x}$ ,  $x_i$  and  $y_i$  exist, but both  $\{x_i\}_{i=1}^\infty$  and  $\{y_i\}_{i=1}^\infty$  do not have finite limits. Again,  $\{(i, 0)\}_{i=1}^\infty$  is a Palais-Smale sequence that does not converge.

It is possible that  $\{x_i\}_{i=1}^\infty$  and  $\{y_i\}_{i=1}^\infty$  have limit points but not a common limit point. To see this, consider the example  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$f(x) = \begin{cases} x & \text{if } x \leq -1 \\ -1 & \text{if } -1 \leq x \leq 1 \\ -x & \text{if } x \geq 1. \end{cases}$$

The set  $\text{lev}_{\leq -1} f$  is path-connected, but the set  $\text{cl}(\text{lev}_{< -1} f)$  is not path-connected. Any point in the set  $(\text{lev}_{\leq -1} f) \setminus \text{cl}(\text{lev}_{< -1} f) = (-1, 1)$  is a local minimum, and hence a critical point.

### 3. A LOCALLY SUPERLINEARLY CONVERGENT ALGORITHM

In this section, we propose a locally superlinearly convergent algorithm for the mountain pass problem for smooth critical points in  $\mathbb{R}^n$ . For this section, we take  $X = \mathbb{R}^n$ . Like Algorithm 2.1 earlier, we keep track of only two points in the space  $\mathbb{R}^n$  instead of a path. Our fast locally convergent algorithm does not require one to calculate the Hessian. Furthermore, we maintain upper and lower bounds that converge superlinearly to the critical value. The numerical performance of this method will be illustrated in Section 8.

In Algorithm 3.1 below, we can assume that the endpoints  $x_0$  and  $y_0$  satisfy  $f(x_0) = f(y_0)$ . Otherwise, if  $f(x_0) < f(y_0)$  say, replace  $x_0$  by the point  $x'_0$  closest to  $x_0$  on the line segment  $[x_0, y_0]$  such that  $f(x'_0) = f(y_0)$ .

**Algorithm 3.1.** (*Fast local level set algorithm*) Find saddle point between points  $x_0$  and  $y_0$  for  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Assume that the objective of the mountain pass problem between  $x_0$  and  $y_0$  is greater than  $f(x_0)$ , and  $f(x_0) = f(y_0)$ . Let  $U$  be a convex set containing  $x_0$  and  $y_0$ .

- (1) Given points  $x_i$  and  $y_i$ , find  $z_i$  as follows:
  - (a) Replace  $x_i$  and  $y_i$  by  $\tilde{x}_i$  and  $\tilde{y}_i$ , where  $\tilde{x}_i$  and  $\tilde{y}_i$  are minimizers of the problem

$$\begin{aligned} \min_{x, y} \quad & |x - y| \\ \text{s.t.} \quad & x \text{ in same component as } x_i \text{ in } (\text{lev}_{\leq f(x_i)} f) \cap U \\ & y \text{ in same component as } y_i \text{ in } (\text{lev}_{\leq f(x_i)} f) \cap U \end{aligned}$$

- (b) Find a minimizer of  $f$  on  $L_i \cap U$ , say  $z_i$ . Here  $L_i$  is the affine space orthogonal to  $x_i - y_i$  passing through  $\frac{1}{2}(x_i + y_i)$ .
- (2) Find the point furthest away from  $x_i$  on the line segment  $[x_i, z_i]$ , which we call  $x_{i+1}$ , such that  $f(x) \leq f(z_i)$  for all  $x$  in the line segment  $[x_i, x_{i+1}]$ . Do the same to find  $y_{i+1}$ .
- (3) Increase  $i$ , repeat steps 1 and 2 until  $|x_i - y_i|$  is small, or if the value  $M_i - f(z_i)$ , where  $M_i := \max_{x \in [x_i, y_i]} f(x)$ , is small.
- (4) If an actual path is desired, take a path  $p_i : [0, 1] \rightarrow X$  lying in  $\text{lev}_{\leq M_i} f$  connecting the points

$$x_0, x_1, \dots, x_{i-2}, x_{i-1}, x_i, y_i, y_{i-1}, y_{i-2}, \dots, y_1, y_0.$$

As we will see in Propositions 4.3 and 5.4, a unique minimizing pair  $(\tilde{x}_i, \tilde{y}_i)$  in step 1(a) exists under added conditions. Furthermore, Proposition 4.5 implies that a unique minimizer of  $f$  on  $L_i \cap U$  exists under added conditions in step 1(b).

To motivate step 1(b), consider any path from  $x_i$  to  $y_i$  in  $U$  that lies wholly in  $U$ . Such a path has to pass through some point of  $L_i \cap U$ , so the maximum value of  $f$  on the path is at least the minimum of  $f$  on  $L_i \cap U$ .

Step 1(a) is analogous to step 2 of Algorithm 2.1. Algorithm 3.1 can be seen as an improvement Algorithm 2.1: The bisection algorithm in Algorithm 2.1 gives us a reliable way of finding the critical point, and step 1(b) in Algorithm 3.1 reduces the distance between the components of the level sets as fast as possible.

In practice, step 1(a) is difficult, and is performed only when the algorithm runs into difficulties. In fact, this step was not performed in our numerical experiments in Section 8. However, we can construct simple functions for which the affine space  $L_i$  does not separate the two components containing  $x_i$  and  $y_i$  in  $(\text{lev}_{\leq f(x_i)} f) \cap U$  in step 1(b) if step 1(a) were not performed.

In the minimum distance problem in step 1(a), notice that if  $f$  is  $\mathcal{C}^1$  and the gradients of  $f$  at a pair of points are nonzero and do not point in opposite directions, then in principle we can perturb the points along paths that decrease the distance between them while not increasing their function values. Of course, a good approximation of a minimizing pair may be hard to compute in practice: existing path-based algorithms for finding mountain passes face analogous computational challenges. One may employ the heuristic in Remark 5.7 for this problem.

In step 2, continuity of  $f$  and  $p$  tells us that  $f(x_{i+1}) = f(z_i)$ . We shall see in Theorem 4.8 that under added conditions,  $\{f(x_i)\}_i$  is an increasing sequence that converges to the critical value  $f(\bar{x})$ . Furthermore, Propositions 4.5 and 5.3 state that under added conditions,  $\{M_i\}_i$  are upper bounds on  $f(\bar{x})$  that converge  $R$ -superlinearly to  $f(\bar{x})$ , where  $R$ -superlinear convergence is defined as follows.

**Definition 3.2.** A sequence in  $\mathbb{R}$  converges *R-superlinearly* to zero if its absolute value is bounded by a superlinearly convergent sequence.

#### 4. SUPERLINEAR CONVERGENCE OF THE LOCAL ALGORITHM

When  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a quadratic whose Hessian has one negative eigenvalue and  $n - 1$  positive eigenvalues, Algorithm 3.1 converges to the critical point in one step. One might expect that if  $f$  is  $\mathcal{C}^2$ , then Algorithm 3.1 converges quickly. In this section, we will prove Theorem 4.8 on the superlinear convergence of Algorithm 3.1.

Recall that the *Morse index* of a critical point is the maximum dimension of a subspace on which the Hessian is negative definite, and a critical point is *nondegenerate* if its Hessian is invertible, and degenerate otherwise. In the smooth finite dimensional case, the Morse index equals the number of negative eigenvalues of the Hessian. If a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mathcal{C}^2$  in a neighborhood of a nondegenerate critical point  $\bar{x}$  of Morse index 1, we can readily make the following assumptions.

**Assumption 4.1.** Assume that  $\bar{x} = \mathbf{0}$  and  $f(\mathbf{0}) = 0$ , and the Hessian  $H = H(\mathbf{0})$  is a diagonal matrix with entries  $a_1, a_2, \dots, a_{n-1}, a_n$  in decreasing order, of which  $a_n$  is negative and  $a_{n-1}$  is the smallest positive eigenvalue.

Another assumption that we will use quite often in this section and the next is on the local approximation of  $f$  near  $\mathbf{0}$ .



**Assumption 4.2.** For  $\delta \in (0, \min\{a_{n-1}, -a_n\})$ , assume  $\theta > 0$  is small enough so that

$$\left| f(x) - \sum_{j=1}^n a_j x^2(j) \right| \leq \delta |x|^2 \text{ for all } x \in \mathbb{B}(\mathbf{0}, \theta).$$

This particular choice of  $\theta$  gives a region  $\mathbb{B}(\mathbf{0}, \theta)$  where Figure 4.1 is valid. We shall use  $\mathring{\mathbb{B}}$  to denote the open ball.

Here is our first result on step 1(a) of Algorithm 3.1.

**Proposition 4.3.** Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mathcal{C}^2$ , and  $\bar{x}$  is a nondegenerate critical point of Morse index 1 such that  $f(\bar{x}) = c$ . If  $\theta > 0$  is sufficiently small, then for any  $\epsilon > 0$  (depending on  $\theta$ ) sufficiently small,

- (1)  $(\text{lev}_{\leq c-\epsilon} f) \cap \mathring{\mathbb{B}}(\bar{x}, \theta)$  has exactly two path connected components, and
- (2) There is a pair  $(\tilde{x}, \tilde{y})$ , where  $\tilde{x}$  and  $\tilde{y}$  lie in distinct components of  $(\text{lev}_{\leq c-\epsilon} f) \cap \mathring{\mathbb{B}}(\bar{x}, \theta)$ , such that  $|\tilde{x} - \tilde{y}|$  is the distance between the two components in  $(\text{lev}_{\leq c-\epsilon} f) \cap \mathring{\mathbb{B}}(\bar{x}, \theta)$ .

*Proof.* Suppose that Assumption 4.1 holds. Choose some  $\delta \in (0, \min\{a_{n-1}, -a_n\})$  and a corresponding  $\theta > 0$  such that Assumption 4.2 holds. A simple bound on  $f(x)$  on  $\mathbb{B}(\mathbf{0}, \theta)$  is therefore:

$$(4.1) \quad \sum_{j=1}^n (a_j - \delta) x^2(j) \leq f(x) \leq \sum_{j=1}^n (a_j + \delta) x^2(j).$$

So if  $\epsilon$  is small enough, the level set  $S := \text{lev}_{\leq -\epsilon} f$  satisfies

$$S_+ \cap \mathbb{B}(\mathbf{0}, \theta) \subset S \cap \mathbb{B}(\mathbf{0}, \theta) \subset S_- \cap \mathbb{B}(\mathbf{0}, \theta),$$

where

$$\begin{aligned} S_+ &:= \left\{ x \mid \sum_{j=1}^n (a_j + \delta) x^2(j) \leq -\epsilon \right\}, \\ S_- &:= \left\{ x \mid \sum_{j=1}^n (a_j - \delta) x^2(j) \leq -\epsilon \right\}, \end{aligned}$$

and  $S_+ \cap \mathbb{B}(\mathbf{0}, \theta)$  is nonempty. Figure 4.1 shows a two-dimensional cross section of the sets  $S_+$  and  $S_-$  through the critical point  $\mathbf{0}$  and the closest points between components in  $S_+$  and  $S_-$ .

**Step 1: Calculate variables in Figure 4.1.**

The two points in distinct components of  $S_+$  closest to each other are the points  $(\mathbf{0}, \pm \sqrt{\frac{\epsilon}{-a_n - \delta}})$ , and one easily calculates the values of  $b$  and  $c$  (which are the distances between  $\mathbf{0}$  and  $S_-$ , and that of  $\mathbf{0}$  and  $S_+$  respectively) in the diagram to be  $\sqrt{\frac{\epsilon}{-a_n + \delta}}$  and  $\sqrt{\frac{\epsilon}{-a_n - \delta}}$ . Thus the distance between the two components of  $S$  is at most  $2\sqrt{\frac{\epsilon}{-a_n - \delta}}$ . The points in  $S$  that minimize the distance between the components must lie in two cylinders  $C_1$  and  $C_2$  defined by

$$(4.2) \quad \begin{aligned} C_1 &:= \mathbb{B}^{n-1}(\mathbf{0}, a) \times [b - 2c, -b] \subset \mathbb{R}^{n-1} \times \mathbb{R}, \\ C_2 &:= \mathbb{B}^{n-1}(\mathbf{0}, a) \times [b, 2c - b] \subset \mathbb{R}^{n-1} \times \mathbb{R}, \end{aligned}$$



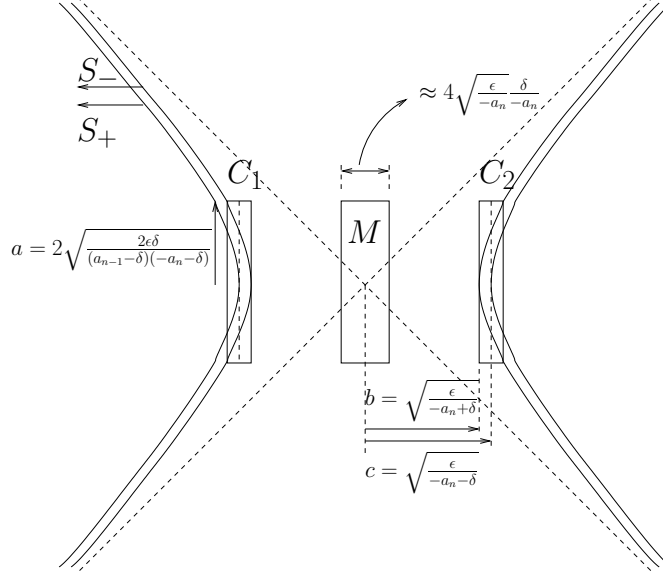


FIGURE 4.1. Local structure of saddle point.

for some  $a > 0$ . In other words,  $C_1$  and  $C_2$  are cylinders with spherical base of radius  $a$  such that

$$(S_- \setminus S_+) \cap (\mathbb{R}^{n-1} \times [b - 2c, 2c - b]) \cap \mathbb{B}(\mathbf{0}, \theta) \subset C_1 \cup C_2.$$

They are represented as the left and right rectangles in Figure 4.1.

We now find a value of  $a$ . We can let  $x(n) = 2c - b$ , and we need

$$\begin{aligned} \sum_{j=1}^{n-1} (a_j - \delta)x^2(j) + (a_n - \delta)x^2(n) &\leq -\epsilon \\ \Rightarrow \sum_{j=1}^{n-1} (a_j - \delta)x^2(j) + (a_n - \delta) \left( 2\sqrt{\frac{\epsilon}{-a_n - \delta}} - \sqrt{\frac{\epsilon}{-a_n + \delta}} \right)^2 &\leq -\epsilon. \end{aligned}$$

Continuing the arithmetic gives

$$\begin{aligned} &\sum_{j=1}^{n-1} (a_j - \delta)x^2(j) \\ &\leq \epsilon \left( -1 - (a_n - \delta) \left( \frac{4}{-a_n - \delta} + \frac{1}{-a_n + \delta} - \frac{4}{\sqrt{-a_n - \delta}\sqrt{-a_n + \delta}} \right) \right) \\ &\leq \epsilon \left( -1 - (a_n - \delta) \left( \frac{4}{-a_n - \delta} + \frac{1}{-a_n + \delta} - \frac{4}{-a_n + \delta} \right) \right) \\ &= \frac{8\epsilon\delta}{-a_n - \delta}. \end{aligned}$$

The radius is maximized when  $x(1) = x(2) = \dots = x(n-2) = 0$  and  $x(n-1) = 2\sqrt{\frac{2\epsilon\delta}{(a_{n-1}-\delta)(-a_n-\delta)}}$ , which gives our value of  $a$ .

**Step 2:**  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$  has exactly two components if  $\epsilon$  is small enough.

Note that  $(\text{lev}_{\leq -\epsilon} f) \cap \mathbb{B}(\mathbf{0}, \theta)$  does not intersect the subspace  $L' := \{x \mid x(n) = 0\}$ , since  $f(x) \geq 0$  for all  $x \in L' \cap \mathbb{B}(\mathbf{0}, \theta)$ . We proceed to show that

$$U_{<} := \{x \mid x(n) < 0\} \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$$

contains exactly one path connected component if  $\epsilon$  is small enough. A similar statement for  $U_{>}$  defined in a similar way will allow us to conclude that  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$  has exactly two components.

Consider two points  $v_1, v_2$  in  $(\text{lev}_{\leq -\epsilon} f) \cap U_{<}$ . We want to find a path connecting  $v_1$  and  $v_2$  and contained in  $(\text{lev}_{\leq -\epsilon} f) \cap U_{<}$ . We may assume that  $v_1(n) \leq v_2(n) < 0$ . By the continuity of the Hessian, assume that  $\theta$  is small enough so that for all  $x \in \mathbb{B}(\mathbf{0}, \theta)$ , the top left principal submatrix of  $H(x)$  corresponding to the first  $n-1$  elements is positive definite. Consider the subspace  $L'(\alpha) := \{x \mid x(n) = \alpha\}$ . The positive definiteness of the submatrix of  $H(x)$  on  $\mathbb{B}(\mathbf{0}, \theta)$  tells us that  $f$  is strictly convex on  $\mathbb{B}(\mathbf{0}, \theta) \cap L'(\alpha)$ .

If  $v_1(n) = v_2(n)$ , then the line segment connecting  $v_1$  and  $v_2$  lies in  $(\text{lev}_{\leq -\epsilon} f) \cap L'(v_1(n)) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$  by the convexity of  $f$  on  $L'(v_1(n)) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ . Otherwise, assume that  $v_1(n) < v_2(n)$ .

Here is a lemma that we will need for the proof.

**Lemma 4.4.** *Suppose Assumption 4.1 holds. We can reduce  $\theta > 0$  and  $\delta > 0$  if necessary so that Assumption 4.2 is satisfied, and the  $n$ th component of  $\nabla f(x)$  is positive for all  $x \in (\text{lev}_{\leq 0} f) \cap \mathbb{B}(\mathbf{0}, \theta) \cap \{x \mid x(n) < 0\}$ .*

*Proof.* We first define  $\tilde{S}_-$  by

$$\tilde{S}_- := \{x \mid (a_{n-1} - \delta) \sum_{j=1}^{n-1} x^2(j) + (a_n - \delta)x^2(n) \leq 0\}.$$

It is clear that  $(a_{n-1} - \delta) \sum_{j=1}^{n-1} x^2(j) + (a_n - \delta)x^2(n) \leq f(x)$  for all  $x \in \mathbb{B}(\mathbf{0}, \theta)$ , so  $(\text{lev}_{\leq 0} f) \cap \mathbb{B}(\mathbf{0}, \theta) \subset \tilde{S}_- \cap \mathbb{B}(\mathbf{0}, \theta)$ .

We now use the expansion  $\nabla f(x) = H(\mathbf{0})x + o(|x|)$ , and prove that the  $n$ th component of  $\nabla f(x)$  is negative for all  $x \in \tilde{S}_- \cap \mathbb{B}(\mathbf{0}, \theta) \cap \{x \mid x(n) < 0\}$ . We can reduce  $\theta$  so that  $|\nabla f(x) - H(\mathbf{0})x| < \delta|x|$  for all  $x \in \mathbb{B}(\mathbf{0}, \theta)$ . Note that if  $x \in \tilde{S}_-$ , then

$$\begin{aligned} (a_{n-1} - \delta) \sum_{j=1}^{n-1} x^2(j) + (a_n - \delta)x^2(n) &\leq 0 \\ \Rightarrow (a_{n-1} - \delta)|x|^2 + (a_n - a_{n-1})x^2(n) &\leq 0 \\ \Rightarrow |x| &\leq \sqrt{\frac{a_{n-1} - a_n}{a_{n-1} - \delta}} (-x(n)). \end{aligned}$$

The  $n$ th component of  $\nabla f(x)$  is bounded from below by

$$a_n x(n) - \delta|x| \leq a_n x(n) + \delta \sqrt{\frac{a_{n-1} - a_n}{a_{n-1} - \delta}} x(n).$$

Provided that  $\delta$  is small enough, the term above is positive since  $x(n) < 0$ .  $\square$

We now return to show that there is a path connecting  $v_1$  and  $v_2$ . Note that  $S_+ \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta) \cap \{x \mid x(n) < 0\}$  is a convex set. (To see this, note that  $S_+ \cap \{x \mid x(n) < 0\}$  can be rotated so that it is the epigraph of a convex function.) Since  $S_+ \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta) \subset (\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ , the open line segment connecting the points  $(\mathbf{0}, -\theta), (\mathbf{0}, -c) \in \mathbb{R}^{n-1} \times \mathbb{R}$  lies in  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ . If  $-\theta < v_1(n) < v_2(n) \leq -c$ , the piecewise linear path connecting  $v_2$  to  $(\mathbf{0}, v_2(n))$  to  $(\mathbf{0}, v_1(n))$  to  $v_1$  lies in  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ .

In the case when  $v_2(n) > -c$ , we see that  $v_2$  must lie in  $C_1$ . Lemma 4.4 tells us that the line segment joining  $v_2$  and  $v_2 + (\mathbf{0}, -c - v_2(n))$  lies in  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ . This allows us to find a path connecting  $v_2$  to  $v_1$ .

**Step 3:  $\tilde{x}$  and  $\tilde{y}$  lie in  $\mathring{\mathbb{B}}(\mathbf{0}, \theta)$ .**

The points  $\tilde{x}$  and  $\tilde{y}$  must lie in  $C_1$  and  $C_2$  respectively, and both  $C_1$  and  $C_2$  lie in  $\mathring{\mathbb{B}}(\mathbf{0}, \theta)$  if  $\epsilon$  is small enough. Therefore, we can minimize over the compact sets  $(\text{lev}_{\leq -\epsilon} f) \cap C_1$  and  $(\text{lev}_{\leq -\epsilon} f) \cap C_2$ , which tells us that a minimizing pair  $(\tilde{x}, \tilde{y})$  exist.  $\square$

In fact, under the assumptions of Proposition 4.3,  $\tilde{x}$  and  $\tilde{y}$  are unique, but all we need in the proof of Proposition 4.5 below is that  $\tilde{x}$  and  $\tilde{y}$  lie in the sets  $C_1$  and  $C_2$  defined by (4.2) respectively and represented as rectangles in Figure 4.1. We defer the proof of uniqueness to Proposition 5.4.

Our next result is on a bound for possible locations of  $z_i$  in step 1(b).

**Proposition 4.5.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mathcal{C}^2$ , and  $\bar{x}$  is a nondegenerate critical point of Morse index 1 such that  $f(\bar{x}) = c$ . If  $\theta$  is small enough, then for all small  $\epsilon > 0$  (depending on  $\theta$ ),*

- (1) *Two closest points of the two components of  $(\text{lev}_{\leq c-\epsilon} f) \cap \mathring{\mathbb{B}}(\bar{x}, \theta)$ , say  $\tilde{x}$  and  $\tilde{y}$ , exist,*
- (2) *For any such points  $\tilde{x}$  and  $\tilde{y}$ ,  $f$  is strictly convex on  $L \cap \mathring{\mathbb{B}}(\bar{x}, \theta)$ , where  $L$  is the orthogonal bisector of  $\tilde{x}$  and  $\tilde{y}$ , and*
- (3)  *$f$  has a unique minimizer on  $L \cap \mathring{\mathbb{B}}(\bar{x}, \theta)$ . Furthermore,  $\min_{L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)} f \leq f(\bar{x}) \leq \max_{[\tilde{x}, \tilde{y}]} f$ .*

*Proof.* Suppose that Assumption 4.1 holds, and choose  $\delta \in (0, \min\{a_{n-1}, -a_n\})$ . Suppose that  $\theta > 0$  is small enough such that Assumption 4.2 holds. Throughout this proof, we assume all vectors accented with a hat ' $\wedge$ ' are of Euclidean length 1. It is clear that  $f(\tilde{x}) = f(\tilde{y}) = -\epsilon$ . Point (1) of the result comes from Proposition 4.3. We first prove the following lemma.

**Lemma 4.6.** *Suppose Assumptions 4.1 and 4.2 hold. If  $\theta > 0$  is small enough, then for all small  $\epsilon > 0$  (depending on  $\theta$ ), two closest points of the two components of  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ , say  $\tilde{x}$  and  $\tilde{y}$ , exist. Let  $L$  be the perpendicular bisector of  $\tilde{x}$  and  $\tilde{y}$ . Then*

$$(\text{lev}_{\leq 0} f) \cap L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta) \subset \mathbb{B}^{n-1} \left( \mathbf{0}, \alpha \sqrt{\frac{(-a_n + \delta)}{(a_{n-1} - \delta)}} \right) \times (-\alpha, \alpha),$$

$$\text{where } \alpha = \delta \sqrt{\frac{\epsilon}{-a_n}} \left( \frac{8}{a_{n-1}} + \frac{2}{-a_n} \right) + o(\delta).$$

*Proof.* By Proposition 4.3, the points  $\tilde{x}$  and  $\tilde{y}$  must exist. We proceed to prove the rest of Lemma 4.6.

**Step 1: Calculate remaining values in Figure 4.1.**

We calculated the values of  $a$ ,  $b$  and  $c$  in step 2 of the proof of Proposition 4.3, and we proceed to calculate the rest of the variables in Figure 4.1. The middle rectangle in Figure 4.1 represents the possible locations of midpoints of points in  $C_1$  and  $C_2$ , and is a cylinder as well. We call this set  $M$ . The radius of this cylinder is the same as that of  $C_1$  and  $C_2$ , and the width of this cylinder is  $4(c - b)$ , which gives an  $o(\delta)$  approximation

$$\begin{aligned}
 4(c - b) &= 4 \left( \sqrt{\frac{\epsilon}{-a_n - \delta}} - \sqrt{\frac{\epsilon}{-a_n + \delta}} \right) \\
 &= 4 \sqrt{\frac{-a_n \epsilon}{(-a_n - \delta)(-a_n + \delta)}} \left( \sqrt{1 + \frac{\delta}{-a_n}} - \sqrt{1 - \frac{\delta}{-a_n}} \right) \\
 &= 4 \sqrt{\frac{\epsilon}{-a_n}} \left( \left( 1 + \frac{\delta}{-2a_n} \right) - \left( 1 - \frac{\delta}{-2a_n} \right) \right) + o(\delta) \\
 &= 4 \sqrt{\frac{\epsilon}{-a_n}} \frac{\delta}{-a_n} + o(\delta).
 \end{aligned}$$

These calculations suffice for the calculations in step 2 of this proof.

**Step 2: Set up optimization problem for bound on  $(\text{lev}_{\leq 0} f) \cap L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ .**

From the values of  $a$  and  $b$  calculated previously, we deduce that a vector  $c_2 - c_1$ , with  $c_i \in C_i$ , can be scaled so that it is of the form  $(\gamma \frac{a}{b} \hat{\mathbf{v}}_1, 1)$ , where  $\hat{\mathbf{v}}_1 \in \mathbb{R}^{n-1}$  is of norm 1 and  $0 \leq \gamma \leq 1$ . (i.e., the norm corresponding to the first  $n - 1$  coordinates is at most  $\frac{a}{b}$ .) These are possible normals for  $L$ , the perpendicular bisector of  $\tilde{x}$  and  $\tilde{y}$ . The formula for  $\frac{a}{b}$  is

$$\begin{aligned}
 \frac{a}{b} &= 2 \sqrt{\frac{2\epsilon\delta}{(a_{n-1} - \delta)(-a_n - \delta)}} \div \sqrt{\frac{\epsilon}{-a_n + \delta}} \\
 &= 2 \sqrt{\frac{2\delta(-a_n + \delta)}{(a_{n-1} - \delta)(-a_n - \delta)}}.
 \end{aligned}$$

So we can represent a normal of the affine space  $L$  as

$$(4.3) \quad \left( 2\gamma_1 \sqrt{\frac{2\delta(-a_n + \delta)}{(a_{n-1} - \delta)(-a_n - \delta)}} \hat{\mathbf{v}}_1, 1 \right) \text{ for some } 0 \leq \gamma_1 \leq 1.$$

We now proceed to bound the minimum of  $f$  on all possible perpendicular bisectors of  $c_1$  and  $c_2$  within  $\mathring{\mathbb{B}}(\mathbf{0}, \theta)$ , where  $c_1 \in C_1$  and  $c_2 \in C_2$ . We find the largest value of  $\alpha$  such that

- there is a point of the form  $(\mathbf{v}_2, \alpha)$  lying in  $\tilde{S}_-$ , where

$$\tilde{S}_- := \{x \mid (a_{n-1} - \delta) \sum_{j=1}^{n-1} x^2(j) + (a_n - \delta)x^2(n) \leq 0\} \subset \mathbb{R}^{n-1} \times \mathbb{R}.$$

- $(\mathbf{v}_2, \alpha) \in \tilde{L}$  for some affine space  $\tilde{L}$  passing through a point  $p \in M$  and having a normal vector of the form in Formula (4.3).

The set  $\tilde{S}_-$  is the same as that defined in the proof of Lemma 4.4. Note that  $\tilde{S}_- \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta) \supset (\text{lev}_{\leq 0} f) \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ , and this largest value of  $\alpha$  is an upper bound on the absolute value of the  $n$ th coordinate of elements in  $(\text{lev}_{\leq 0} f) \cap L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ .

**Step 3: Solving for  $\alpha$ .**

For a point  $(\mathbf{v}_2, \alpha) \in \tilde{S}_-$ , where  $\mathbf{v}_2 = (x(1), x(2), \dots, x(n-1)) \in \mathbb{R}^{n-1}$ , we have

$$\begin{aligned} (a_{n-1} - \delta) \sum_{j=1}^{n-1} x^2(j) + (a_n - \delta) \alpha^2 &\leq 0. \\ \Rightarrow |\mathbf{v}_2|^2 &= \sum_{j=1}^{n-1} x^2(j) \\ &\leq \frac{(-a_n + \delta)}{(a_{n-1} - \delta)} \alpha^2. \\ \Rightarrow |\mathbf{v}_2| &\leq \sqrt{\frac{(-a_n + \delta)}{(a_{n-1} - \delta)}} \alpha. \end{aligned}$$

Therefore, we can write  $(\mathbf{v}_2, \alpha)$  as

$$(4.4) \quad \left( \gamma_2 \sqrt{\frac{(-a_n + \delta)}{(a_{n-1} - \delta)}} \alpha \hat{\mathbf{v}}_2, \alpha \right),$$

where  $\hat{\mathbf{v}}_2 \in \mathbb{R}^{n-1}$  is a vector of unit norm, and  $0 \leq \gamma_2 \leq 1$ . We can assume that  $p$  has coordinates

$$\left( 2\gamma_3 \sqrt{\frac{2\epsilon\delta}{(a_{n-1} - \delta)(-a_n - \delta)}} \hat{\mathbf{v}}_3, 2\gamma_4 \sqrt{\frac{\epsilon}{-a_n - a_n}} \frac{\delta}{-a_n - a_n} + o(\delta) \right),$$

where  $\hat{\mathbf{v}}_3 \in \mathbb{R}^{n-1}$  is some vector of unit norm, and  $0 \leq \gamma_3, \gamma_4 \leq 1$ . Note that the  $n$ th component is half the width of  $M$ . Hence a possible tangent on  $\tilde{L}$  is

$$\left( \gamma_1 \sqrt{\frac{(-a_n + \delta)}{(a_{n-1} - \delta)}} \alpha \hat{\mathbf{v}}_2, \alpha \right) - \left( 2\gamma_3 \sqrt{\frac{2\epsilon\delta}{(a_{n-1} - \delta)(-a_n - \delta)}} \hat{\mathbf{v}}_3, 2\gamma_4 \sqrt{\frac{\epsilon}{-a_n - a_n}} \frac{\delta}{-a_n - a_n} + o(\delta) \right).$$

To simplify notation, note that we only require an  $O(\delta)$  approximation of  $\alpha$ , we can take the terms like  $-a_n + \delta$  and  $-a_n - \delta$  to be  $-a_n + O(\delta)$  and so on. The dot product of the above vector and the normal of the affine space  $L$  calculated in Formula (4.3) must be zero, which after some simplification gives:

$$\begin{aligned} &\left( \left( \gamma_2 \sqrt{\frac{-a_n}{a_{n-1}}} + O(\delta) \right) \alpha \hat{\mathbf{v}}_2 - \left( 2\gamma_3 \sqrt{\frac{2\epsilon\delta}{a_{n-1}(-a_n)}} + O(\delta^{3/2}) \right) \hat{\mathbf{v}}_3 \right. \\ &\quad \left. , \alpha - \left( 2\gamma_4 \sqrt{\frac{\epsilon}{-a_n - a_n}} \frac{\delta}{-a_n - a_n} + o(\delta) \right) \right) \cdot \left( \left( 2\gamma_1 \sqrt{\frac{2\delta}{a_{n-1}}} + O(\delta^{3/2}) \right) \hat{\mathbf{v}}_1, 1 \right) = 0. \end{aligned}$$

At this point, we remind the reader that the  $O(\delta^k)$  terms mean that there exists some  $K > 0$  such that if  $\delta$  were small enough, we can find terms  $t_1$  to  $t_3$  such that  $|t_i| < K\delta^k$  and the formula above is satisfied by  $t_i$  in place of the  $O(\delta^k)$  terms. Further arithmetic gives

$$\begin{aligned} &4\gamma_1\gamma_3 \sqrt{\frac{2\delta}{a_{n-1}}} \sqrt{\frac{2\epsilon\delta}{a_{n-1}(-a_n)}} (\hat{\mathbf{v}}_3 \cdot \hat{\mathbf{v}}_1) + 2\gamma_4 \sqrt{\frac{\epsilon}{-a_n - a_n}} \frac{\delta}{-a_n - a_n} + o(\delta) \\ &= \alpha \left( 1 + 2\gamma_1\gamma_2 \sqrt{\frac{2\delta}{a_{n-1}}} \sqrt{\frac{-a_n}{a_{n-1}}} (\hat{\mathbf{v}}_2 \cdot \hat{\mathbf{v}}_1) + o(\delta^{3/2}) \right) \\ &= \alpha(1 + O(\sqrt{\delta})) \end{aligned}$$

To find an upper bound for  $\alpha$ , it is clear that we should take  $\gamma_1 = \gamma_3 = \gamma_4 = 1$  and  $\hat{\mathbf{v}}_3 \cdot \hat{\mathbf{v}}_1 = 1$ . The  $O(\sqrt{\delta})$  term is superfluous, and this simplifies to give

$$(4.5) \quad \alpha \leq \delta \sqrt{\frac{\epsilon}{-a_n}} \left( \frac{8}{a_{n-1}} + \frac{2}{-a_n} \right) + o(\delta).$$

We could find the minimum possible value of  $\alpha$  by these same series of steps and show that the absolute value would be bounded above by the same bound. This ends the proof of Lemma 4.6.  $\square$

It is clear that the minimum value of  $f$  on  $L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$  is at most 0, since  $L$  intersects the axis corresponding to the  $n$ th coordinate and  $f$  is nonpositive there. Therefore the set  $(\text{lev}_{\leq 0} f) \cap L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$  is nonempty, and  $f$  has a local minimizer on  $L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ .

We now state and prove our second lemma that will conclude the proof of Proposition 4.5.

**Lemma 4.7.** *Let  $L$  be the perpendicular bisector of  $\tilde{x}$  and  $\tilde{y}$  as defined in point (1) of Proposition 4.5 with  $\tilde{x} = \mathbf{0}$ . If  $\delta$  and  $\theta$  are small enough satisfying Assumptions 4.1 and 4.2, then  $f|_{L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)}$  is strictly convex.*

*Proof.* The lineality space of  $L$ , written as  $\text{lin}(L)$ , is the space of vectors orthogonal to  $\tilde{x} - \tilde{y}$ . We can infer from Formula (4.3) that  $\tilde{x} - \tilde{y}$  is a scalar multiple of a vector of the form  $(w, 1)$ , where  $w \in \mathbb{R}^{n-1}$  satisfies  $|w| \rightarrow \mathbf{0}$  as  $\delta \rightarrow 0$ . We consider a vector  $v \in \text{lin}(L)$  orthogonal to  $(w, 1)$  that can be scaled so that  $v = (\tilde{w}, 1)$ , where  $(w, 1) \cdot (\tilde{w}, 1) = 0$ , which gives  $w \cdot \tilde{w} = -1$ . The Cauchy Schwarz inequality gives us

$$\begin{aligned} |\tilde{w}| |w| &\geq |\tilde{w} \cdot w| \\ &= 1 \\ \Rightarrow |\tilde{w}| &\geq |w|^{-1}. \end{aligned}$$

So

$$\begin{aligned} \frac{v^\top H(p)v}{v^\top v} &= \frac{v^\top H(\mathbf{0})v}{v^\top v} + \frac{v^\top (H(p) - H(\mathbf{0}))v}{v^\top v} \\ &= \frac{\sum_{j=1}^{n-1} a_j v^2(j) + a_n}{\sum_{j=1}^{n-1} v^2(j) + 1} + \frac{v^\top (H(p) - H(\mathbf{0}))v}{v^\top v} \\ &\geq \underbrace{\frac{a_{n-1} \sum_{j=1}^{n-1} v^2(j) + a_n}{\sum_{j=1}^{n-1} v^2(j) + 1}}_{(1)} + \underbrace{\frac{v^\top (H(p) - H(\mathbf{0}))v}{v^\top v}}_{(2)}. \end{aligned}$$

Since  $\sum_{j=1}^{n-1} v^2(j) = |\tilde{w}|^2 \rightarrow \infty$  as  $|w| \rightarrow 0$ , the limit of term (1) is  $a_{n-1}$ , so there is an open set  $\mathbb{B}(\mathbf{0}, \theta)$  containing  $\mathbf{0}$  such that  $\frac{v^\top H(p)v}{v^\top v} > \frac{1}{2}a_{n-1}$  for all  $v \in \text{lin}(L) \cap \{x \mid x(n) = 1\}$  and  $p \in \mathbb{B}(\mathbf{0}, \theta)$ . By the continuity of the Hessian, we may reduce  $\theta$  if necessary so that  $\|H(p) - H(\mathbf{0})\| < \frac{1}{2}a_{n-1}$  for all  $p \in \mathbb{B}(\mathbf{0}, \theta)$ . Thus  $\frac{v^\top H(p)v}{v^\top v} > 0$  for all  $p \in \mathbb{B}(\mathbf{0}, \theta)$  and  $v \in \text{lin}(L) \cap \{x \mid x(n) = 1\}$  if  $\delta$  is small enough.

The vectors of the form  $v = (\tilde{w}, 0)$  do not present additional difficulties as the corresponding term (1) is at least  $a_{n-1}$ . This proves that the Hessian  $H(p)$  restricted to  $\text{lin}(L)$  is positive definite, and hence the strict convexity of  $f$  on  $L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$ .  $\square$

Since  $f$  has a local minimizer in  $L \cap \mathring{\mathbb{B}}(\mathbf{0}, \theta)$  and is strictly convex there, we have (2) and the first part of part (3). The inequality  $f(\bar{x}) \leq \max_{[\tilde{x}, \tilde{y}]} f$  follows easily from the fact that the line segment  $[\tilde{x}, \tilde{y}]$  intersects the set  $\{x \mid x(n) = 0\}$ , on which  $f$  is nonnegative.  $\square$

Here is our theorem on the convergence of Algorithm 3.1.

**Theorem 4.8.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mathcal{C}^2$  in a neighborhood of a nondegenerate critical point  $\bar{x}$  of Morse index 1. If  $\theta > 0$  is sufficiently small and  $x_0$  and  $y_0$  are chosen such that*

- (a)  $x_0$  and  $y_0$  lie in the two different components of  $(\text{lev}_{\leq f(x_0)} f) \cap \mathring{\mathbb{B}}(\bar{x}, \theta)$ ,
- (b)  $f(x_0) = f(y_0) < f(\bar{x})$ ,

then Algorithm 3.1 with  $U = \mathring{\mathbb{B}}(\bar{x}, \theta)$  generates a sequence of iterates  $\{\tilde{x}_i\}_i$  and  $\{\tilde{y}_i\}_i$  lying in  $\mathring{\mathbb{B}}(\bar{x}, \theta)$  such that the function values  $\{f(\tilde{x}_i)\}_i$  and  $\{f(\tilde{y}_i)\}_i$  converge to  $f(\bar{x})$  superlinearly, and the iterates  $\{\tilde{x}_i\}_i$  and  $\{\tilde{y}_i\}_i$  converge to  $\bar{x}$  superlinearly.

*Proof.* As usual, suppose Assumption 4.1 holds, and  $\delta$  and  $\theta$  are chosen so that Assumption 4.2 holds.

**Step 1: Linear convergence of  $f(\tilde{x}_i)$  to critical value  $f(\bar{x})$ .**

Let  $\epsilon = f(\tilde{x}_i)$ . The next iterate  $x_{i+1}$  satisfies  $f(x_{i+1}) = f(z_i)$ , and is bounded from below by

$$f(x_{i+1}) \geq (a_n - \delta)\alpha^2 = -\epsilon\delta^2 \left( \frac{8}{a_{n-1}} + \frac{2}{-a_n} \right)^2 + o(\delta^2),$$

where  $\alpha$  is the value calculated in Lemma 4.6. The ratio between the previous function value and the next function value is at most

$$\rho(\delta) := \delta^2 \left( \frac{8}{a_{n-1}} + \frac{2}{-a_n} \right)^2 + o(\delta^2).$$

This ratio goes to 0 as  $\delta \searrow 0$ , so we can choose some  $\delta$  small enough so that  $\rho < \frac{1}{2}$ . We can choose  $\theta$  corresponding to the value of  $\delta$  satisfying Assumption 4.2. This shows that the convergence to 0 of the function values  $f(\tilde{x}_{i+1}) = f(x_{i+1})$  in Algorithm 3.1 is linear provided  $x_0$  and  $y_0$  lie in  $\mathbb{B}(\mathbf{0}, \theta)$  and  $\epsilon$  is small enough by Proposition 4.3. We can reduce  $\theta$  if necessary so that  $f(x) \geq -\epsilon$  for all  $x \in \mathbb{B}(\mathbf{0}, \theta)$ , so the condition on  $\epsilon$  does not present difficulties.

**Step 2: Superlinear convergence of  $f(\tilde{x}_i)$  to critical value  $f(\bar{x})$ .**

Choose a sequence  $\{\delta_k\}_k$  so that  $\delta_k \searrow 0$  monotonically. Corresponding to  $\delta_k$ , we can choose  $\theta_k$  satisfying Assumption 4.2. Since  $\{\tilde{x}_i\}_i$  and  $\{\tilde{y}_i\}_i$  converge to  $\mathbf{0}$ , for any  $k \in \mathbb{Z}_+$ , we can find some  $i^* \in \mathbb{Z}_+$  so that the cylinders  $C_1$  and  $C_2$  constructed in Figure 4.1 corresponding to  $\epsilon_i = -f(\tilde{x}_i)$  and  $\delta = \delta_1$  lie wholly in  $\mathbb{B}(\mathbf{0}, \theta_k)$  for all  $i > i^*$ . As remarked in step 3 of the proof of Proposition 4.3,  $\tilde{x}_i$  and  $\tilde{y}_i$  must lie inside  $C_1$  and  $C_2$ , so we can take  $\delta = \delta_k$  for the ratio  $\rho$ . This means that  $\frac{|f(\tilde{x}_{i+1})|}{|f(\tilde{x}_i)|} \leq \rho(\delta_k)$  for all  $i > i^*$ . As  $\rho(\delta) \searrow 0$  as  $\delta \searrow 0$ , this means that we have superlinear convergence of the  $f(\tilde{x}_i)$  to the critical value  $f(\bar{x})$ .

**Step 3: Superlinear convergence of  $\tilde{x}_i$  to the critical point  $\bar{x}$ .**

We now proceed to prove that the distance between the critical point  $\mathbf{0}$  and the iterates decrease superlinearly by calculating the value  $\frac{|\tilde{x}_{i+1}|}{|\tilde{x}_i|}$ , or alternatively  $\frac{|\tilde{x}_{i+1}|^2}{|\tilde{x}_i|^2}$ . The value  $|\tilde{x}_i|$  satisfies  $|\tilde{x}_i|^2 \geq b^2 = \frac{\epsilon}{-a_n + \delta}$ . To find an upper bound for  $|\tilde{x}_{i+1}|^2$ , it is instructive to look at an upper bound for  $|\tilde{x}_i|^2$  first. As can be deduced



from Figure 4.1, an upper bound for  $|\tilde{x}_i|^2$  is the square of the distance between  $\mathbf{0}$  and the furthest point in  $C_1$ , which is

$$\begin{aligned} (2c-b)^2 + a^2 &= (c + (c-b))^2 + a^2 \\ &= \frac{\epsilon}{-a_n - \delta} + 8 \frac{\epsilon \delta}{(-a_n)^2} + \frac{8\epsilon \delta}{(a_{n-1} - \delta)(-a_n - \delta)} + o(\delta). \end{aligned}$$

This means that an upper bound for  $|\tilde{x}_{i+1}|^2$  is

$$\delta^2 \left( \frac{8}{a_{n-1}} + \frac{2}{-a_n} \right)^2 \left( \frac{\epsilon}{-a_n - \delta} + \frac{8\epsilon \delta}{-a_n} \left( \frac{1}{-a_n} + \frac{1}{(a_{n-1} - \delta)} \right) \right) + o(\delta^2).$$

From this point, one easily sees that as  $i \rightarrow \infty$ ,  $\delta \rightarrow 0$ , and  $\frac{|\tilde{x}_{i+1}|^2}{|\tilde{x}_i|^2} \rightarrow 0$ . This gives the superlinear convergence of the distance between the critical point and the iterates  $\tilde{x}_i$  that we seek.  $\square$

## 5. FURTHER PROPERTIES OF THE LOCAL ALGORITHM

In this section, we take note of some interesting properties of Algorithm 3.1. First, we show that it is easy to find  $x_{i+1}$  and  $y_{i+1}$  in step 2 of Algorithm 3.1.

**Proposition 5.1.** *Suppose the conditions in Theorem 4.8 hold. Consider the sequence of iterates  $\{x_i\}_i$  and  $\{y_i\}_i$  generated by Algorithm 3.1. If  $i$  is large enough, then either  $x_{i+1} = z_i$  or  $y_{i+1} = z_i$  in step 2 of Algorithm 3.1.*

*Proof.* Let  $\tilde{p} : [0, 1] \rightarrow \mathbb{R}^n$  denote the piecewise linear path connecting  $x_i$  to  $z_i$  to  $y_i$ . It suffices to prove that along  $\tilde{p}$ , the function  $f$  increases to a maximum, and then decreases. Suppose Assumptions 4.1 and 4.2 hold. The cylinders  $C_1$  and  $C_2$  in Figure 4.1 are loci for  $x_i$  and  $y_i$ . We assume that  $x_i$  lies in  $C_2$  in Figure 4.1. The calculations in (4.4) in Lemma 4.6 tell us that  $z_i$  can be written as

$$\left( \sqrt{\frac{(-a_n + \delta)}{(a_{n-1} - \delta)}} \alpha \lambda_1 \hat{\mathbf{v}}_2, \lambda_2 \alpha \right) \in \mathbb{R}^{n-1} \times \mathbb{R},$$

where  $0 < \lambda_1 < \lambda_2 \leq 1$ ,  $|\hat{\mathbf{v}}_2| = 1$  and  $\alpha = \delta \sqrt{\frac{\epsilon}{-a_n}} \left( \frac{8}{a_{n-1}} + \frac{2}{-a_n} \right) + o(\delta)$  by (4.5). Therefore,  $x_i - z_i$  can be written as

$$\left( \mathbf{v}_1, \sqrt{\frac{\epsilon}{-a_n + \delta}} + o(\sqrt{\delta \epsilon}) \right),$$

where  $\mathbf{v}_1 \in \mathbb{R}^{n-1}$  satisfies

$$\begin{aligned} |\mathbf{v}_1| &\leq \sqrt{\frac{(-a_n + \delta)}{(a_{n-1} - \delta)}} \alpha + a \\ &= O(\sqrt{\epsilon \delta}), \end{aligned}$$

and  $a = \sqrt{\frac{2\epsilon \delta}{(a_{n-1} - \delta)(-a_n - \delta)}}$  is as calculated in the proof of Proposition 4.3. This means that the unit vector with direction  $x_i - z_i$  converges to the  $n$ -th elementary vector as  $\delta \searrow 0$ . By appealing to Hessians as is done in the proof of Lemma 4.7, we see that the function  $f$  is strictly concave in the line segment  $[x_i, z_i]$  if  $i$  is large enough. Similarly,  $f$  is strictly concave in the line segment  $[y_i, z_i]$  if  $i$  is large enough.

Next, we prove that the function  $f$  has only one local maximizer in  $\tilde{p}([0, 1])$ . In the case where  $\nabla f(z_i) = \mathbf{0}$ , the concavity of  $f$  on the line segments  $[x_i, z_i]$  and  $[y_i, z_i]$  tells us that  $z_i$  is the unique maximizer on  $\tilde{p}([0, 1])$ . We now look at the case where  $\nabla f(z_i) \neq \mathbf{0}$ . Since  $z_i$  is the minimizer on a subspace with normal  $x_i - y_i$ ,  $\nabla f(z_i)$  is a (possibly negative) multiple of  $x_i - y_i$ . This means that  $\nabla f(z_i) \cdot (x_i - z_i)$  has a different sign than  $\nabla f(z_i) \cdot (y_i - z_i)$ . In other words, the map  $t \mapsto f(\tilde{p}(t))$  increases then decreases. This concludes the proof of the proposition.  $\square$

*Remark 5.2.* Note that in Algorithm 3.1, all we need in step 1 is a good lower bound of the critical value. We can exploit convexity as proved in Lemma 4.7 and use cutting plane methods to attain a lower bound for  $f$  on  $L_i \cap \mathbb{B}(\bar{x}, \theta)$ .

Recall from Proposition 4.5 that  $M_i$  is a sequence of upper bounds of the critical value  $f(\bar{x})$ . While it is not even clear that  $M_i$  is monotonically decreasing, we can prove the following convergence result on  $M_i$ .

**Proposition 5.3.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mathcal{C}^2$  in a neighborhood of a nondegenerate critical point  $\bar{x}$  of Morse index 1, the neighborhood  $U$  of  $\bar{x}$  and the points  $x_0$  and  $y_0$  are chosen satisfying the conditions in the statement of Theorem 4.8. Then in Algorithm 3.1,  $M_i := \max_{[x_i, y_i]} f$  converges  $R$ -superlinearly to the critical value.*

*Proof.* Suppose Assumption 4.1 holds. An upper bound of the critical value of the saddle point is obtained by finding the maximum along the line segment joining two points in  $C_1$  and  $C_2$ , which is bounded from above by

$$(a_1 + \delta)a^2 = (a_1 + \delta) \frac{8\epsilon\delta}{(a_{n-1} - \delta)(-a_n - \delta)}.$$

A more detailed analysis by using cylinders with ellipsoidal base instead of circular base tell us that the maximum is bounded above by  $\frac{8\epsilon\delta}{(-a_n - \delta)}$  instead. If  $\delta > 0$  is small enough, this value is much smaller than  $-f(x_i) = \epsilon$ . As  $i \rightarrow \infty$ , the estimates  $-f(x_i)$  converge superlinearly to 0 by Theorem 4.8, giving us what we need.  $\square$

Step 1(a) is important in the analysis of Algorithm 3.1. As explained earlier in Section 3, it may be difficult to implement this step. Algorithm 3.1 may run fine without ever performing step 1(a) (see the example in Section 8), but it may need to be performed occasionally in a practical implementation. The following result tells us that under the assumptions we have made so far, this problem is locally a strictly convex problem with a unique solution.

**Proposition 5.4.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mathcal{C}^2$  in a neighborhood of a nondegenerate critical point  $\bar{x}$  of Morse index 1 with critical value  $f(\bar{x}) = c$ . Then if  $\epsilon > 0$  is small enough, there is a convex neighborhood  $U_\epsilon$  of  $\bar{x}$  such that  $(\text{lev}_{\leq c-\epsilon} f) \cap U_\epsilon$  is a union of two disjoint convex sets.*

*Consequently, providing  $\theta$  is sufficiently small, the pair of nearest points guaranteed by Proposition 4.3(2) are unique.*

*Proof.* Suppose Assumptions 4.1 and 4.2 hold. In addition, we further assume that

$$|\nabla f(x) - H(x)| < \delta |x| \text{ for all } x \in \mathring{\mathbb{B}}(\mathbf{0}, \theta).$$

We can choose  $U_\epsilon$  to be the interior of  $\text{conv}(C_1 \cup C_2)$ , where  $C_1$  and  $C_2$  are the cylinders in Figure 4.1 and defined in the proof of Proposition 4.3, but in view of

Theorem 5.6, we shall prove that  $U_\epsilon$  can be chosen to be the bigger set  $\text{conv}(\tilde{C}_1 \cup \tilde{C}_2)$ , where  $\tilde{C}_1$  and  $\tilde{C}_2$  are cylinders defined by

$$\begin{aligned}\tilde{C}_1 &:= \mathbb{B}^{n-1}(\mathbf{0}, \rho) \times [-\beta, -b] \subset \mathbb{R}^{n-1} \times \mathbb{R}, \\ \tilde{C}_2 &:= \mathbb{B}^{n-1}(\mathbf{0}, \rho) \times [b, \beta] \subset \mathbb{R}^{n-1} \times \mathbb{R},\end{aligned}$$

where  $\beta, \rho$  are constants to be determined. We choose  $\beta$  such that

$$\mathbb{B}^{n-1}(\mathbf{0}, a) \times \{\beta\} \subset \text{int}(S_+).$$

In particular,  $\beta$  satisfies

$$\begin{aligned}a^2(a_1 + \delta) + \beta^2(a_n + \delta) &< -\epsilon \\ \Rightarrow \beta^2 &> \frac{1}{-a_n - \delta} (\epsilon + a^2(a_1 + \delta)) \\ &= \frac{\epsilon}{-a_n - \delta} \left( 1 + \frac{8\delta(a_1 + \delta)}{(a_{n-1} - \delta)(-a_n - \delta)} \right)\end{aligned}$$

We choose  $\beta$  to be any value satisfying the above inequality.

Next, we choose  $\rho$  to be the smallest value such that  $S_- \cap (\mathbb{R}^{n-1} \times [-\beta, \beta]) \cap \mathbb{B}(\mathbf{0}, \theta) \subset \tilde{C}_1 \cup \tilde{C}_2$ . This calculation is similar to the calculation of  $a$ , which gives

$$\begin{aligned}(a_{n-1} - \delta)\rho^2 + (a_n - \delta)\beta^2 &= -\epsilon \\ \Rightarrow \rho &= \sqrt{\frac{-\epsilon - (a_n - \delta)\beta^2}{a_{n-1} - \delta}}.\end{aligned}$$

We shall not expand the terms, but remark that  $\beta$  and  $\rho$  are of  $O(\sqrt{\epsilon})$ .

The proof of Proposition 4.3 tells us that  $\text{conv}(\tilde{C}_1 \cup \tilde{C}_2) \cap \text{lev}_{\leq -\epsilon} f$  is a union of the two nonempty sets  $\tilde{C}_1 \cap \text{lev}_{\leq -\epsilon} f$  and  $\tilde{C}_2 \cap \text{lev}_{\leq -\epsilon} f$ . It remains to show that these two sets are strictly convex.

Any point  $x \in \tilde{C}_1$  can be written as

$$x = (\mathbf{x}', x_n),$$

where  $\mathbf{x}' \in \mathbb{R}^{n-1}$  is of norm at most  $\rho$ , and  $-\beta \leq x_n \leq -b$ , where  $\beta$  is as calculated above and  $b = \sqrt{\frac{\epsilon}{-a_n + \delta}}$  as in Figure 4.1. This implies that

$$Hx = (\mathbf{x}'', a_n x_n),$$

where  $\mathbf{x}''$  is of norm at most  $a_1 |\mathbf{x}'|$ . It is clear that as  $\delta \downarrow 0$ , the unit vector in the direction of  $Hx$  converges to  $(\mathbf{0}, 1)$ . This implies that for any  $\kappa_1 > 0$ , there exists some  $\delta > 0$  such that  $\text{unit}(\nabla f(x)) \cdot (\mathbf{0}, 1) \geq 1 - \kappa_1$  for all  $x \in \tilde{C}_1$ . (Note that  $\tilde{C}_1$  depends on  $\delta$ .) Here,  $\text{unit} : \mathbb{R}^n \setminus \{\mathbf{0}\} \rightarrow \mathbb{R}^n$  is the mapping of a nonzero vector to the unit vector pointing in the same direction.

Let  $z_1$  and  $z_2$  be points in  $\tilde{C}_1 \cap (\text{lev}_{\leq -\epsilon} f)$ . Suppose that  $z_1(n) < z_2(n)$ , and let  $\mathbf{v} = (\mathbf{v}_1, v_2) \in \mathbb{R}^{n-1} \times \mathbb{R}$  be a unit vector in the same direction as  $z_2 - z_1$ . We further assume, by reducing  $\theta$  and  $\delta$  as necessary, that  $\|H(x) - H(\mathbf{0})\| < \kappa_2$  for all  $x \in \tilde{C}_1 \cap (\text{lev}_{\leq -\epsilon} f)$ . Suppose  $\kappa_1$  and  $\kappa_2$  are small enough so that  $\sqrt{2\kappa_1} < \sqrt{\frac{a_{n-1} - \kappa_2}{a_{n-1} - a_n}}$ .

Note that  $v_2 \geq 0$ . Either one of these two cases on  $v_2$  must hold. We prove that in both cases, the open line segment  $(z_1, z_2)$  lies in the interior of  $(\text{lev}_{\leq -\epsilon} f) \cap \tilde{C}_1$ .

**Case 1:**  $v_2 > \sqrt{2\kappa_1}$ .

In this case, for all  $x \in \tilde{C}_1$ , we have

$$\begin{aligned}
\mathbf{v} \cdot (\text{unit}(\nabla f(x))) &= \mathbf{v} \cdot (\mathbf{0}, 1) + \mathbf{v} \cdot (\text{unit}(\nabla f(x)) - (\mathbf{0}, 1)) \\
&\geq v_2 - |\mathbf{v}| |\text{unit}(\nabla f(x)) - (\mathbf{0}, 1)| \\
&= v_2 - |\text{unit}(\nabla f(x)) - (\mathbf{0}, 1)| \\
&= v_2 - \sqrt{|\text{unit}(\nabla f(x))|^2 + |(\mathbf{0}, 1)|^2 - 2\text{unit}(\nabla f(x)) \cdot (\mathbf{0}, 1)} \\
&> v_2 - \sqrt{2 - 2(1 - \kappa_1)} \\
&= v_2 - \sqrt{2\kappa_1} \\
&> 0.
\end{aligned}$$

This means that along the line segment  $[z_1, z_2]$ , the function  $f$  is strictly monotone. Therefore, if  $x_1, x_2 \in (\text{lev}_{\leq -\epsilon} f) \cap \tilde{C}_1$ , the open line segment  $(z_1, z_2)$  lies in the interior of  $(\text{lev}_{\leq -\epsilon} f) \cap \tilde{C}_1$ .

**Case 2:**  $v_2 < \sqrt{\frac{a_{n-1} - \kappa_2}{a_{n-1} - a_n}}$ .

Let  $H^u(\mathbf{0})$  denote the diagonal matrix of size  $(n-1) \times (n-1)$  with elements  $a_1, \dots, a_{n-1}$ . We have

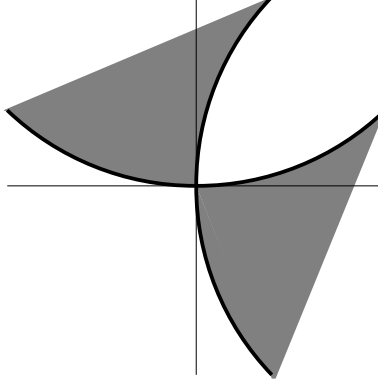
$$\begin{aligned}
\mathbf{v}^\top H(x) \mathbf{v} &= \mathbf{v}^\top H(\mathbf{0}) \mathbf{v} + \mathbf{v}^\top (H(x) - H(\mathbf{0})) \mathbf{v} \\
&> \mathbf{v}_1^\top H^u(\mathbf{0}) \mathbf{v}_1 + a_n v_2^2 - |\mathbf{v}|^2 \|H(x) - H(\mathbf{0})\| \\
&\geq a_{n-1} |\mathbf{v}_2|^2 + a_n v_2^2 - \|H(x) - H(\mathbf{0})\| \\
&> a_{n-1} (1 - v_2^2) + a_n v_2^2 - \kappa_2 \\
&= a_{n-1} + v_2^2 (a_n - a_{n-1}) - \kappa_2 \\
&> a_{n-1} + (\kappa_2 - a_{n-1}) - \kappa_2 \\
&\geq 0
\end{aligned}$$

This means that the function  $f$  is strictly convex along the line segment  $[z_1, z_2]$ , so if  $x_1, x_2 \in (\text{lev}_{\leq -\epsilon} f) \cap \tilde{C}_1$ , the open line segment  $(z_1, z_2)$  lies in the interior of  $(\text{lev}_{\leq -\epsilon} f) \cap \tilde{C}_1$ , concluding the proof of the first part of this result.

To prove the next statement on the uniqueness of the pair of closest points, suppose that  $(\tilde{x}', \tilde{y}')$  and  $(\tilde{x}'', \tilde{y}'')$  are distinct pairs whose distance give the distance between the components of  $(\text{lev}_{\leq -\epsilon} f) \cap \mathbb{B}(\mathbf{0}, \theta)$ , where  $\mathbb{B}(\mathbf{0}, \theta)$  is as stated in Proposition 4.3. If  $\epsilon$  is small enough, then  $\text{conv}(\tilde{C}_1 \cup \tilde{C}_2)$  lies in  $\mathring{\mathbb{B}}(\mathbf{0}, \theta)$ . Then by the strict convexity of the components of  $(\text{lev}_{\leq -\epsilon} f) \cap \text{conv}(\tilde{C}_1 \cup \tilde{C}_2)$ , the pair  $(\frac{1}{2}(\tilde{x}' + \tilde{x}''), \frac{1}{2}(\tilde{y}' + \tilde{y}''))$  lie in the same components, and the distance between this pair of points must be the same as that for the pairs  $(\tilde{x}', \tilde{y}')$  and  $(\tilde{x}'', \tilde{y}'')$ . The closest points in the components of  $[\frac{1}{2}(\tilde{x}' + \tilde{x}''), \frac{1}{2}(\tilde{y}' + \tilde{y}'')] \cap \text{lev}_{\leq -\epsilon} f$  give a smaller distance between the components of  $(\text{lev}_{\leq -\epsilon} f) \cap \mathbb{B}(\mathbf{0}, \theta)$ , which contradicts the optimality of the pairs  $(\tilde{x}', \tilde{y}')$  and  $(\tilde{x}'', \tilde{y}'')$ .  $\square$

Note that in the case of  $\epsilon = 0$ , there may be no neighborhood  $U_0$  of  $\bar{x}$  such that  $U_0 \cap (\text{lev}_{\leq c} f)$  is a union of two convex sets intersecting only at the critical point. We also note that  $U_\epsilon$  depends on  $\epsilon$  in our result above. The following example explains these restrictions.

**Example 5.5.** Consider the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  defined by  $f(x) = (x_2 - x_1^2)(x_1 - x_2^2)$ . The shaded area in Figure 5.1 is a sketch of  $\text{lev}_{\leq 0} f$ .


 FIGURE 5.1.  $\text{lev}_{\leq 0} f$  for  $f(x) = (x_2 - x_1^2)(x_1 - x_2^2)$ 

We now explain that the neighborhood  $U_\epsilon$  defined in Proposition 5.4 must depend on  $\epsilon$  for this example. For any open  $U$  containing  $\mathbf{0}$ , we can always find two points  $p$  and  $q$  in a component of  $(\text{lev}_{< 0} f) \cap U$  such that the line segment  $[p, q]$  does not lie in  $\text{lev}_{< 0} f$ . This implies that the component of  $(\text{lev}_{\leq -\epsilon} f) \cap U$  is not convex if  $0 < \epsilon \leq -\max(f(p), f(q))$ .  $\diamond$

We now take a second look at the problem of minimizing the distance between two components in step 1(a) of Algorithm 3.1. We need to solve the following problem for  $\epsilon > 0$ :

$$\begin{aligned}
 (5.1) \quad & \min_{x, y} \quad |x - y| \\
 & \text{s.t.} \quad x \text{ lies in the same component as } a \text{ in } (\text{lev}_{\leq f(\bar{x}) - \epsilon} f) \cap \mathring{\mathbb{B}}(\bar{x}, \theta) \\
 & \quad y \text{ lies in the same component as } b \text{ in } (\text{lev}_{\leq f(\bar{x}) - \epsilon} f) \cap \mathring{\mathbb{B}}(\bar{x}, \theta).
 \end{aligned}$$

If  $(\tilde{x}, \tilde{y})$  is a pair of local optimizers, then  $\tilde{y}$  is the closest point to the component of  $(\text{lev}_{\leq f(\bar{x}) - \epsilon} f) \cap U$  containing  $\tilde{x}$  and vice versa. This gives us the following optimality conditions:

$$\begin{aligned}
 (5.2) \quad & \nabla f(\tilde{x}) = \kappa_1(\tilde{y} - \tilde{x}), \\
 & \nabla f(\tilde{y}) = \kappa_2(\tilde{x} - \tilde{y}), \\
 & f(\tilde{x}) = f(\bar{x}) - \epsilon \\
 & f(\tilde{y}) = f(\bar{x}) - \epsilon \\
 & \text{for some } \kappa_1, \kappa_2 \geq 0.
 \end{aligned}$$

From Proposition 5.4, we see that given any  $\theta > 0$  sufficiently small, provided that the conditions in Proposition 4.3 hold, the global minimizing pair of (5.1) is unique. Even though convexity is absent, the following theorem shows that the global minimizing pair is, under added conditions, the only pair satisfying the optimality conditions (5.2), showing that there are no other local minimizers of (5.1).

**Theorem 5.6.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mathcal{C}^2$ , and  $\bar{x}$  is a nondegenerate critical point of Morse index 1 such that  $f(\bar{x}) = c$ . If  $\theta > 0$  is sufficiently small, then for any  $\epsilon > 0$  (depending on  $\theta$ ) sufficiently small, the global minimizer of (5.1) is the only pair in  $\mathring{\mathbb{B}}(\bar{x}, \theta) \times \mathring{\mathbb{B}}(\bar{x}, \theta)$  satisfying the optimality conditions (5.2).*

*Proof.* Suppose that Assumption 4.1 holds, and  $\delta$  is chosen small enough so that 4.2 holds. We also assume that  $\theta$  is small enough so that  $|H(x) - H(\mathbf{0})| < \frac{1}{2} \min(a_{n-1}, -a_n)$ . Seeking a contradiction, suppose that  $(\tilde{x}, \tilde{y})$  satisfy the optimality conditions.

We refer to Figure 4.1, and also recall the definitions of the sets  $\tilde{C}_1$  and  $\tilde{C}_2$  in the proof of Proposition 5.4. As proven in Proposition 5.4, the convexity properties of the two level sets in  $(\text{lev}_{\leq f(\tilde{x})-\epsilon} f) \cap \mathring{\mathbb{B}}(\tilde{x}, \theta)$  imply that if  $\tilde{x} \in \tilde{C}_1$ ,  $\tilde{y} \in \tilde{C}_2$  and the optimality conditions are satisfied, then the pair  $(\tilde{x}, \tilde{y})$  is the global minimizing pair.

Consider the case where  $\tilde{x} \notin \tilde{C}_1$ . Either of the two cases hold. We note the asymmetry below in that we check whether  $\tilde{y} \in C_2$  instead of whether  $\tilde{y} \in \tilde{C}_2$ .

**Case 1:**  $\tilde{y} \in C_2$ : In this case, if the first  $n-1$  coordinates of  $\tilde{x}$  are the same as that of  $\tilde{y}$ , then  $\tilde{x}$  lies in the interior of  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\tilde{x}, \theta)$ , which is a contradiction to optimality. Recall that the value of  $\beta$  was chosen such that  $\tilde{y} + (\mathbf{0}, \tilde{x}(n) - \tilde{y}(n))$  lies in  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\tilde{x}, \theta)$ . By the convexity of  $f|_{L'(\tilde{x}(n))}$ , where  $L'(\tilde{x}(n))$  is the affine space  $\{x \mid x(n) = \tilde{x}(n)\}$ , the line segment connecting  $\tilde{x}$  and  $\tilde{y} + (\mathbf{0}, \tilde{x}(n) - \tilde{y}(n))$  lies in  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\tilde{x}, \theta)$ . The distance between  $\tilde{y}$  and points along this line segment decreases (at a linear rate) as one moves away from  $\tilde{x}$ , which again contradicts the assumption that  $(\tilde{x}, \tilde{y})$  satisfy (5.2).

**Case 2:**  $\tilde{y} \notin C_2$ : By the convexity of  $f|_{L'(\tilde{x}(n))}$  and  $f|_{L'(\tilde{y}(n))}$ , the line segments  $[\tilde{y}, \tilde{y} - (\mathbf{0}, \tilde{y}(n))]$  and  $[\tilde{x}, \tilde{x} - (\mathbf{0}, \tilde{x}(n))]$  lie in  $(\text{lev}_{\leq -\epsilon} f) \cap \mathring{\mathbb{B}}(\tilde{x}, \theta)$ . These line segments and the optimality of the pair  $(\tilde{x}, \tilde{y})$  implies that the first  $n-1$  components of  $\tilde{x}$  and  $\tilde{y}$  to be the same. This in turn implies that  $\nabla f(\tilde{x})$  is a positive multiple of  $(\mathbf{0}, 1)$ .

Our proof ends if we show that if  $\theta$  is small enough,  $\nabla f(\tilde{x})$  cannot be a positive multiple of  $(\mathbf{0}, 1)$ . If  $\tilde{x} \notin \tilde{C}_1$ , then  $\tilde{x}(n) < -\beta$ . If  $\tilde{x}$  lies on the boundary of  $\text{lev}_{\leq -\epsilon} f$ , then  $f(\tilde{x}) = -\epsilon$ , and we have

$$\begin{aligned} f(\tilde{x}) &= -\epsilon \\ \sum_{i=1}^n (a_i + \delta) \tilde{x}(i)^2 &\geq -\epsilon \\ (a_1 + \delta) \sum_{i=1}^n \tilde{x}(i)^2 + (a_n - a_1) \tilde{x}(n)^2 &\geq -\epsilon \\ (a_1 + \delta) |\tilde{x}|^2 &\geq (a_1 - a_n) \tilde{x}(n)^2 - \epsilon \\ \frac{|\tilde{x}|^2}{\tilde{x}(n)^2} &\geq \frac{a_1 - a_n - \frac{\epsilon}{\tilde{x}(n)^2}}{a_1 + \delta} \\ &\geq 1 + \frac{-a_n - \delta - \frac{\epsilon}{\beta^2}}{a_1 + \delta} \end{aligned}$$

Upon expansion of the term  $\beta^2$  in the expression in the final line, we see that  $\frac{|\tilde{x}|^2}{\tilde{x}(n)^2}$  is bounded from below by a constant independent of  $\epsilon$  and greater than 1. Since  $f$  is  $\mathcal{C}^2$ , the set

$$\{x \mid \nabla f(x) \text{ is a multiple of } (\mathbf{0}, 1)\} \cap \mathbb{B}(\mathbf{0}, \theta)$$

is a manifold, whose tangent at the origin is the line spanned by  $(\mathbf{0}, 1)$ . This implies that if  $\theta$  is small enough, then  $\tilde{x} \notin \tilde{C}_1$  and  $\tilde{x}$  lying on the boundary of

$\text{lev}_{\leq -\epsilon} f$  implies that  $\nabla f(\tilde{x})$  cannot be a multiple of  $(0, 1)$ . We have the required contradiction.  $\square$

*Remark 5.7.* We now describe a heuristic to approximate a pair of closest points iteratively between the components of  $(\text{lev}_{\leq c-\epsilon} f) \cap U$ . For two points  $x'$  and  $y'$  that approximate  $\tilde{x}_i$  and  $\tilde{y}_i$ , we can find local minimizers of  $f$  on the affine spaces orthogonal to  $x' - y'$  that pass through  $x'$  and  $y'$  respectively, say  $x^*$ ,  $y^*$ , and then find the closest points in the two components of  $(\text{lev}_{\leq c-\epsilon} f) \cap [x^*, y^*]$ , where  $[x^*, y^*]$  is the line segment connecting  $x^*$  and  $y^*$ . This heuristic is particularly practical in the case of Wilkinson problem, as we illuminate in Sections 7 and 8.

## 6. SADDLE POINTS AND CRITICALITY PROPERTIES

We have seen that Algorithm 2.1 allows us to find saddle points of mountain type. In this section, we first prove an equivalent definition of a saddle point based on paths connecting two points. Then we prove that saddle points are critical points in the metric sense and in the nonsmooth sense.

In the following equivalent condition for saddle points, we say that a path  $p : [0, 1] \rightarrow X$  connects  $a$  and  $b$  if  $p(0) = a$  and  $p(1) = b$ , and it is contained in  $U \subset X$  if  $p([0, 1]) \subset U$ . The maximum value of the path  $p$  is defined as  $\max_t f \circ p(t)$ .

**Proposition 6.1.** *Let  $(X, d)$  be a metric space. For a continuous function  $f : X \rightarrow \mathbb{R}$ ,  $\bar{x}$  is a saddle point of mountain pass type if and only if there exists an open neighborhood  $U$  and two points  $a, b \in (\text{lev}_{< l} f) \cap U$  such that*

- (a) *The maximum value of any path connecting  $a$  and  $b$  contained in  $U$  is at least  $f(\bar{x})$ , and*
- (b) *for all  $\epsilon > 0$ , there exists  $\delta, \theta \in (0, \epsilon)$  and a path  $p_\epsilon$  connecting  $a$  and  $b$  contained in  $U$  such that the maximum value of  $p_\epsilon$  is at most  $f(\bar{x}) + \epsilon$ , and  $(\text{lev}_{\geq f(\bar{x})-\theta} f) \cap p_\epsilon([0, 1]) \subset \mathbb{B}(\bar{x}, \delta)$ .*

*Proof.* We first prove that the conditions (a) and (b) above imply that  $\bar{x}$  is a saddle point. Let  $A$  and  $B$  be the path connected components of  $\text{lev}_{< f(\bar{x})} f \cap U$  containing  $a$  and  $b$  respectively. For any  $\epsilon > 0$ , the condition  $(\text{lev}_{\geq f(\bar{x})-\theta} f) \cap p_\epsilon([0, 1]) \subset \mathbb{B}(\bar{x}, \delta)$  tells us that we can find points  $x_\epsilon \in A$  and  $y_\epsilon \in B$  such that  $d(\bar{x}, x_\epsilon) < \delta < \epsilon$  and  $d(\bar{x}, y_\epsilon) < \epsilon$ . For a sequence  $\epsilon_i \searrow 0$ , we set  $x_i = x_{\epsilon_i}$  and  $y_i = y_{\epsilon_i}$ . This shows that  $\bar{x}$  lies in both the closure of  $A$  and that of  $B$ , and hence  $\bar{x}$  is a saddle point.

Next, we prove the converse. Suppose that  $\bar{x}$  is a saddle point, with  $U$  being a neighborhood of  $\bar{x}$ , and the sets  $A$  and  $B$  are two path components of  $(\text{lev}_{< f(\bar{x})} f) \cap U$  whose closures contain  $\bar{x}$ . For any  $\epsilon > 0$ , we can find some  $\delta \in (0, \epsilon)$  such that  $d(x, \bar{x}) < \delta$  implies  $|f(x) - f(\bar{x})| < \epsilon$ . There are two points  $x_\epsilon \in A$  and  $y_\epsilon \in B$  such that  $d(x_\epsilon, \bar{x}) < \delta$  and  $d(y_\epsilon, \bar{x}) < \delta$ .

Let  $a$  and  $b$  be any two points in the sets  $A$  and  $B$  respectively. There is a path connecting  $a$  to  $x_\epsilon$  contained in  $\text{lev}_{< f(\bar{x})} f \cap U$ , say  $p_a$ , and we can similarly find a path  $p_b$  connecting  $y_\epsilon$  to  $b$  contained in  $\text{lev}_{< f(\bar{x})} f \cap U$ . The maximum values on both paths  $p_a$  and  $p_b$  are less than  $f(\bar{x})$ , so there is some  $\theta \in (0, \epsilon)$  such that both maximum values are bounded above by  $f(\bar{x}) - \theta$ . Choose a path  $p'_\epsilon$  to be the line segment connecting  $x_a$  and  $y_b$  contained in  $\mathbb{B}(\bar{x}, \delta)$ . The path  $p_\epsilon$  formed by the concatenation of the paths  $p_a$ ,  $p'_\epsilon$  and  $p_b$  satisfies condition (b). Condition (a) is easily seen to be satisfied, and hence we are done.  $\square$



Ideally, we want to improve condition (b) in Proposition 6.1 so that  $\bar{x}$  is the maximum point on some mountain pass connecting  $a$  and  $b$ . We shall see in Example 6.3 that saddle points in general need not have this property. A simple finite dimensional condition on the function  $f$  so that this happens is semi-algebraicity. A set in  $\mathbb{R}^n$  is *semi-algebraic* if it is a union of finitely many sets defined by finitely many polynomial inequalities, and a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is *semi-algebraic* if its graph  $\{(x, y) \in \mathbb{R}^n \times \mathbb{R} \mid y = f(x)\}$  is a semi-algebraic set. Semi-algebraic objects remove much of the oscillatory behavior that typically does not appear in applications, and form a large class of objects that appear in applications. We will appeal to semi-algebraic geometry for only the next result, and we refer readers interested in the general theory of semi-algebraic functions (and more generally, that of o-minimal structures and tame topology, under which Proposition 6.2 also holds) to [7, 16, 15, 20].

**Proposition 6.2.** *In the case where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is semi-algebraic, condition (b) in Proposition 6.1 can be replaced with*

(b') *There is a path connecting  $a$  and  $b$  contained in  $U$  along which the unique maximizer is  $\bar{x}$ .*

*Proof.* It is clear that (b') is a stronger condition than (b), so we prove that if  $f$  is semi-algebraic, then (b') holds. Suppose  $\bar{x}$  is a saddle point of mountain pass type. Let  $U$  be an open neighborhood of  $\bar{x}$ , and sets  $A$  and  $B$  be two components of  $(\text{lev}_{<f(\bar{x})}f) \cap U$  whose closures contain  $\bar{x}$ . Choose points  $a \in A$  and  $b \in B$ . It is clear that  $A$  and  $B$  are semi-algebraic (see for example [15, Section 3.2]). By the curve selection lemma (see for example [15, Section 3.1]), there is a path  $p_a$  connecting  $a$  and  $\bar{x}$  such that  $p_a(1) = \bar{x}$ , and  $p_a([0, 1)) \subset A$ . Similarly, we can find a path  $p_b$  connecting  $\bar{x}$  and  $b$  such that  $p_b(0) = \bar{x}$  and  $p_b((0, 1]) \subset B$ . The concatenation of  $p_a$  and  $p_b$  gives us what we need.  $\square$

In the absence of semi-algebraicity, the following example illustrates that a saddle point need not satisfy condition (b').

**Example 6.3.** We define  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  through Figure 6.1. There are 2 shapes in the positive quadrant the figure: a blue “comb”  $C$  wrapping around a brown “sun”  $S$ . The closure of  $C$  contains the origin  $\mathbf{0}$  (the intersection of the horizontal and vertical axis).

We can define a continuous  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  so that  $f$  is negative on  $C \cup (-C)$  and positive on  $(S \cup (-S)) \setminus \{\mathbf{0}\}$  and  $\{(x, y) \mid xy < 0\}$ , and extend  $f$  continuously to all of  $\mathbb{R}^2$  using the Tietze extension theorem. It is clear that  $\mathbf{0}$  is a saddle point, and the sets  $A, B \subset \text{lev}_{<0}f$  whose closures contain  $\mathbf{0}$  can be taken to be the path connected components containing  $C$  and  $(-C)$  respectively. But the origin  $\mathbf{0}$  does not satisfy condition (b').

Our next step is to establish the relation between saddle points and criticality in metric spaces. We recall the following definitions in metric critical point theory from [17, 23, 25].

**Definition 6.4.** Let  $(X, d)$  be a metric space. We call the point  $x$  *Morse regular* for the function  $f : X \rightarrow \mathbb{R}$  if, for some numbers  $\gamma, \sigma > 0$ , there is a continuous function

$$\phi : \mathbb{B}(x, \gamma) \times [0, \gamma] \rightarrow X$$

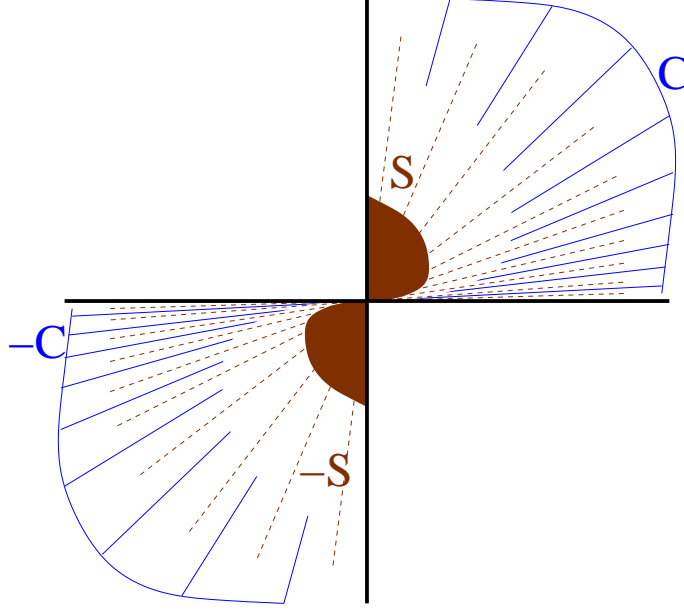


FIGURE 6.1. Illustration of saddle point in Example 6.3.

such that all points  $u \in \mathbb{B}(x, \gamma)$  and  $t \in [0, \gamma]$  satisfy the inequality

$$f(\phi(x, t)) \leq f(x) - \sigma t,$$

and that  $\phi(\cdot, 0)$  is the identity map. The point  $x$  is *Morse critical* if it is not Morse regular.

If there is some  $\kappa > 0$  and such a function  $\phi$  that also satisfies the inequality

$$d(\phi(x, t), x) \leq \kappa t,$$

then we call  $x$  *deformationally regular*. The point  $x$  is *deformationally critical* if it is not deformationally regular.

We now relate saddle points to Morse critical and deformationally critical points.

**Proposition 6.5.** *For a function  $f : X \rightarrow \mathbb{R}$  defined on a metric space  $X$ ,  $\bar{x}$  is a saddle point of mountain pass type implies that  $\bar{x}$  is deformationally critical. If in addition, either  $X = \mathbb{R}^n$  or condition (b') in Proposition 6.2 holds, then  $\bar{x}$  is Morse critical.*

*Proof.* Let  $U$  be an open neighborhood of  $\bar{x}$  as defined in Definition 1.2, and let  $A$  and  $B$  be two distinct components of  $(\text{lev}_{<f(\bar{x})} f) \cap U$  which contain  $\bar{x}$  in their closures. The proofs of all three results by contradiction are similar. For convenience, we label the following three assumptions as follows, and prove that they all lead to the contradiction that  $A$  and  $B$  cannot be distinct path components in  $U$ .

- (D)  $\bar{x}$  is deformationally regular.
- ( $M_{\mathbb{R}^n}$ )  $\bar{x}$  is Morse regular, and  $X = \mathbb{R}^n$ .
- ( $M_{b'}$ )  $\bar{x}$  is Morse regular, and condition (b') in Proposition 6.2 holds.

Suppose condition  $(M_{\mathbb{R}^n})$  holds. Let  $\gamma, \sigma > 0$  and  $\phi : \mathbb{B}(\bar{x}, \gamma) \times [0, \gamma] \rightarrow X$  satisfy the properties of Morse regularity given in Definition 6.4. We can assume that  $\gamma$  is small enough so that  $\mathbb{B}(\bar{x}, \gamma) \subset U$ . By the continuity of  $\phi$  and the compactness of  $\mathbb{B}(\bar{x}, \gamma)$ , there is some  $\gamma' > 0$  such that  $\mathbb{B}(\bar{x}, \gamma) \times [0, \gamma'] \subset \phi^{-1}(U)$ .

Next, suppose condition  $(D)$  holds. Let  $\gamma, \sigma, \kappa > 0$  and  $\phi : \mathbb{B}(\bar{x}, \gamma) \times [0, \gamma] \rightarrow X$  satisfy the properties given in Definition 6.4 on deformation regularity. We can assume  $\gamma > 0$  is small enough and choose  $\gamma' > 0$  so that  $\mathbb{B}(\bar{x}, \gamma + \gamma'\kappa) \subset U$ . The conditions on  $\phi$  imply that  $\phi(\mathbb{B}(\bar{x}, \gamma) \times [0, \gamma']) \subset \mathbb{B}(\bar{x}, \gamma + \gamma'\kappa) \subset U$ , which in turn imply that  $\mathbb{B}(\bar{x}, \gamma) \times [0, \gamma'] \subset \phi^{-1}(U)$ .

Here is the next argument common to both conditions  $(D)$  and  $(M_{\mathbb{R}^n})$ . By the characterization of saddle points in Proposition 6.1, we can find  $\theta$  and  $\delta$  satisfying the condition in Proposition 6.1(b) with  $\theta, \delta \leq \min(\frac{1}{2}\gamma'\sigma, \gamma)$ . This gives us  $\mathbb{B}(\bar{x}, \delta) \subset \mathbb{B}(\bar{x}, \gamma) \subset U$  in particular. We can glean from the proof of Proposition 6.1 that we can find two points  $a_\delta \in A \cap \mathbb{B}(\bar{x}, \delta)$  and  $b_\delta \in B \cap \mathbb{B}(\bar{x}, \delta)$  and a path  $p' : [0, 1] \rightarrow X$  connecting  $a_\delta$  and  $b_\delta$  contained in  $\mathbb{B}(\bar{x}, \delta)$  with maximum value at most  $f(\bar{x}) + \min(\frac{1}{2}\gamma'\sigma, \gamma)$ . The functions values  $f(a_\delta)$  and  $f(b_\delta)$  satisfy  $f(a_\delta), f(b_\delta) \leq f(\bar{x}) - \theta$ . The condition  $\mathbb{B}(\bar{x}, \gamma) \times [0, \gamma'] \subset \phi^{-1}(U)$  implies that  $p'([0, 1]) \times [0, \gamma'] \subset \phi^{-1}(U)$ .

If condition  $(M_{b'})$  holds, then for any  $\delta > 0$ , we can find a path  $p' : [0, 1] \rightarrow X$  connecting two points  $a_\delta \in A \cap \mathbb{B}(\bar{x}, \delta)$  and  $b_\delta \in B \cap \mathbb{B}(\bar{x}, \delta)$  contained in  $\mathbb{B}(\bar{x}, \delta)$  with maximum value at most  $f(\bar{x})$ . There is also some  $\theta > 0$  such that  $f(a_\delta), f(b_\delta) < f(\bar{x}) - \theta$ . Let  $\gamma, \sigma > 0$  and  $\phi : \mathbb{B}(\bar{x}, \gamma) \times [0, \gamma] \rightarrow X$  be such that they satisfy the properties of Morse regularity. By the compactness of  $p'([0, 1])$ , we can find some  $\gamma' > 0$  such that  $p'([0, 1]) \times [0, \gamma'] \subset \phi^{-1}(U)$ .

To conclude the proof for all three cases, consider the path  $\bar{p} : [0, 3] \rightarrow X$  defined by

$$\bar{p}(t) = \begin{cases} \phi(a_\delta, \gamma't) & \text{for } 0 \leq t \leq 1 \\ \phi(p'(t-1), \gamma') & \text{for } 1 \leq t \leq 2 \\ \phi(b_\delta, \gamma'(3-t)) & \text{for } 2 \leq t \leq 3. \end{cases}$$

This path connects  $a_\delta$  and  $b_\delta$ , is contained in  $U$  and has maximum value at most  $\max(f(\bar{x}) - \theta, f(\bar{x}) - \frac{1}{2}\gamma'\sigma)$ , which is less than  $f(\bar{x})$ . This implies that  $A$  and  $B$  cannot be distinct path connected components of  $(\text{lev}_{<f(\bar{x})}f) \cap U$ , which establishes the contradiction in all three cases.  $\square$

We now move on to discuss how saddle points and deformationally critical points relate to nonsmooth critical points. Here is the definition of Clarke critical points.

**Definition 6.6.** [14, Section 2.1] Let  $X$  be a Banach space. Suppose  $f : X \rightarrow \mathbb{R}$  is locally Lipschitz. The *Clarke generalized directional derivative* of  $f$  at  $x$  in the direction  $v \in X$  is defined by

$$f^\circ(x; v) = \limsup_{t \searrow 0, y \rightarrow x} \frac{f(y + tv) - f(y)}{t},$$

where  $y \in X$  and  $t$  is a positive scalar. The *Clarke subdifferential* of  $f$  at  $x$ , denoted by  $\partial_C f(x)$ , is the convex subset of the dual space  $X^*$  given by

$$\{\zeta \in X^* \mid f^\circ(x; v) \geq \langle \zeta, v \rangle \text{ for all } v \in X\}.$$

The point  $x$  is a *Clarke (nonsmooth) critical point* if  $\mathbf{0} \in \partial_C f(x)$ . Here,  $\langle \cdot, \cdot \rangle : X^* \times X \rightarrow \mathbb{R}$  defined by  $\langle \zeta, v \rangle := \zeta(v)$  is the dual relation.

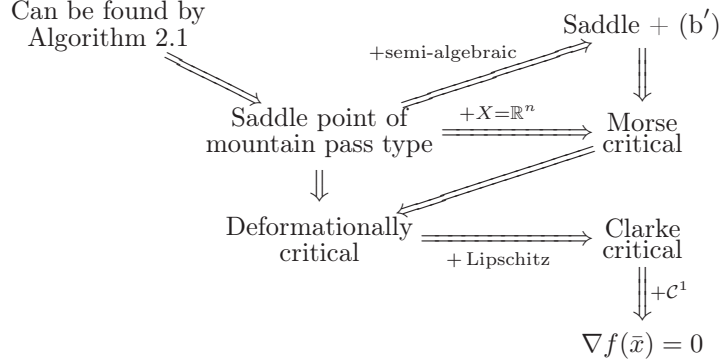


FIGURE 6.2. Different types of critical points

For the particular case of  $\mathcal{C}^1$  functions,  $\partial_C f(x) = \{\nabla f(x)\}$ . Therefore a critical point of a smooth function (i.e., a point  $x$  that satisfies  $\nabla f(x) = \mathbf{0}$ ) is also a Clarke critical point. From the definitions above, it is clear that an equivalent definition of a Clarke critical point is  $f^\circ(x; v) \geq 0$  for all  $v \in X$ . This property allows us to deduce Clarke criticality without appealing to the dual space  $X^*$ .

Clarke (nonsmooth) critical points of  $f$  are of interest in, for example, partial differential equations with discontinuous nonlinearities. Critical point existence theorems for nonsmooth functions first appeared in [12, 37]. For the problem of finding nonsmooth critical points numerically, we are only aware of [44].

The following result is well-known, and we include its proof for completeness.

**Proposition 6.7.** *Let  $X$  be a Banach space and  $f : X \rightarrow \mathbb{R}$  be locally Lipschitz at  $\bar{x}$ . If  $\bar{x}$  is deformationally critical, then it is Clarke critical.*

*Proof.* We prove the contrapositive instead. If the point  $\bar{x}$  is not Clarke critical, there exists a unit vector  $v \in X$  such that

$$\limsup_{t \searrow 0, y \rightarrow \bar{x}} \frac{f(y + tv) - f(y)}{t} < 0.$$

Now defining  $\phi(x, t) = x - tv$  satisfies the conditions for deformation regularity.  $\square$

To conclude, Figure 6.2 summarizes the relationship between saddle points and the different types of critical points.

## 7. WILKINSON'S PROBLEM: BACKGROUND

In Section 8, we will apply Algorithm 3.1 to attempt to solve the Wilkinson problem, while we give a background of the Wilkinson problem in this section. We first define the Wilkinson problem.

**Definition 7.1.** Given a matrix  $A \in \mathbb{R}^{n \times n}$ , the *Wilkinson distance* of the matrix  $A$  is the distance of the matrix  $A$  to the nearest matrix with repeated eigenvalues. The problem of finding the Wilkinson distance is the *Wilkinson problem*.

Though not cited explicitly, as noted by [1], the Wilkinson problem can be traced back to [41, pp. 90-93]. See [2, 10, 28] for more references, and in particular, [2] and the discussion in the beginning of [10, Section 3].

It is well-known that eigenvalues vary in a Lipschitz manner if and only if they do not coincide. In fact, eigenvalues are differentiable in the entries of the matrix when they are distinct. Hence, as discussed by Demmel [18], the Wilkinson distance is a natural condition measure for accurate eigenvalue computation. The Wilkinson distance is also important because of its connections with the stability of eigendecompositions of matrices. To our knowledge, no fast and reliable numerical method for computing the Wilkinson distance is known.

The  $\epsilon$ -pseudospectrum  $\Lambda_\epsilon(A) \subset \mathbb{C}$  of  $A$  is defined as the set

$$\begin{aligned}\Lambda_\epsilon(A) &:= \{z \mid \exists E \text{ s.t. } \|E\| \leq \epsilon \text{ and } z \text{ is an eigenvalue of } A + E\} \\ &= \left\{z \mid \|(A - zI)^{-1}\|^{-1} \leq \epsilon\right\} \\ &= \{z \mid \underline{\sigma}(A - zI) \leq \epsilon\},\end{aligned}$$

where  $\underline{\sigma}(A - zI)$  is the smallest singular value of  $A - zI$ . The function  $z \mapsto (A - zI)^{-1}$  is sometimes referred to as the resolvent function, whose (Clarke) critical points are referred to as *resolvent critical points*. To simplify notation, define  $\underline{\sigma}_A : \mathbb{C} \rightarrow \mathbb{R}_+$  by

$$\begin{aligned}\underline{\sigma}_A(z) &:= \underline{\sigma}(A - zI) \\ &= \text{smallest singular value of } (A - zI).\end{aligned}$$

For more on pseudospectra, we refer the reader to [40].

It is well known that each component of the  $\epsilon$ -pseudospectrum  $\Lambda_\epsilon(A)$  contains at least one eigenvalue. If  $\epsilon$  is small enough,  $\Lambda_\epsilon(A)$  has  $n$  components, each containing an eigenvalue. Alam and Bora [1] proved the following result on the Wilkinson distance.

**Theorem 7.2.** [1] *Let  $\bar{\epsilon}$  be the smallest  $\epsilon$  for which  $\Lambda_\epsilon(A)$  contains  $n - 1$  or fewer components. Then  $\bar{\epsilon}$  is the Wilkinson distance for  $A$ .*

*For any pair of distinct eigenvalues of  $A$ , say  $\{z_1, z_2\}$ , let the objective of the mountain pass problem with function  $\underline{\sigma}_A$  and the two chosen eigenvalues as end-points be  $v(z_1, z_2)$ . The value  $\bar{\epsilon}$  is also equal to*

$$(7.1) \quad \min\{v(z_1, z_2) \mid z_1 \text{ and } z_2 \text{ are distinct eigenvalues of } A\}.$$

Two components of  $\Lambda_\epsilon(A)$  would coalesce when  $\epsilon \uparrow \bar{\epsilon}$ , and the point at which two components coalesce can be used to construct the matrix closest to  $A$  with repeated eigenvalues. Equivalently, the point of coalescence of the two components is also the highest point on an optimal mountain pass for the function  $\underline{\sigma}_A$  between the corresponding eigenvalues. We use Algorithm 3.1 to find such points of coalescence, which are resolvent critical points.

We should remark that solving for  $v(z_1, z_2)$  is equivalent to solving a global mountain pass problem, which is difficult. Also, the problem of finding the eigenvalue pair  $\{z_1, z_2\}$  that minimizes (7.1) is potentially difficult. In Section 8, we focus only on finding a critical point of mountain pass type between two chosen eigenvalues  $z_1$  and  $z_2$ . Fortunately, this strategy often succeeds in obtaining the Wilkinson distance in our experiments in Section 8.

We should note that other approaches for the Wilkinson problem include [2], which uses a Newton type method for the same local problem, and [30].

## 8. WILKINSON'S PROBLEM: IMPLEMENTATION AND NUMERICAL RESULTS

We first use a convenient fast heuristic to estimate which pseudospectral components first coalesce as  $\epsilon$  increases from zero, as follows. We construct the Voronoi diagram corresponding to the spectrum, and then minimize the function  $\underline{\sigma}_A : \mathbb{C} \rightarrow \mathbb{R}$  over all the line segments in the diagram (a fast computation, as discussed in the comments on Step 1(b) below). We then concentrate on the pair of eigenvalues separated by the line segment containing the minimizer. This is illustrated in Example 8.1 below.

We describe implementation issues of Algorithm 3.1.

**Step 1(a):** Approximately minimizing the distance between a pair of points in distinct components seem challenging in practice, as we discussed briefly in Section 3. In the case of pseudospectral components, we have the advantage that computing the intersection between any circle and the pseudospectral boundary is an easy eigenvalue computation [31]. This observation can be used to check optimality conditions or algorithm design for step 1(a). We note that in our numerical implementation, step 1(a) is never actually performed.

**Step 1(b):** Finding the global minimizer in step 1(b) of Algorithm 3.1 is easy in this case. Byers [11] proved that  $\epsilon$  is a singular value of  $A - (x + iy)I$  if and only if  $iy$  is an eigenvalue of

$$\begin{pmatrix} x - A^* & -\epsilon I \\ \epsilon I & A - x \end{pmatrix}.$$

Using Byer's observation, Boyd and Balakrishnan [9] devised a globally convergent and locally quadratic convergent method for the minimization problem over  $\mathbb{R}$  of  $y \mapsto \underline{\sigma}_A(x + iy)$ . We can easily amend these observations to calculate the minimum of  $\underline{\sigma}_A(x + iy)$  over a line segment efficiently by noticing that if  $|z| = 1$ , then

$$\underline{\sigma}_A(x + iy) = \underline{\sigma}(A - (x + iy)I) = \underline{\sigma}(z(A - (x + iy)I)).$$

**Example 8.1.** We apply our mountain pass algorithm on the matrix

$$A = \begin{pmatrix} .461 + .650i & .006 + .625i & & & \\ & .457 + .983i & .297 + .733i & & \\ & & .451 + .553i & .049 + .376i & \\ & & & .412 + .400i & .693 + .010i \\ & & & & .902 + .199i \end{pmatrix}$$

The results of the numerical algorithm are presented in Table 1, and plots using EigTooL [43] are presented in Figure 8.1. We tried many random examples of bidiagonal matrices taking entries in the square  $\{x + iy \mid 0 \leq x, y \leq 1\}$  of the same form as  $A$ . The convergence to a critical point in this example is representative of the typical behavior we encountered.

In Figure 8.1, the top left picture shows that the first step in the Voronoi diagram method identifies the pseudospectral components corresponding to the eigenvalues  $0.461 + 0.650i$  and  $0.451 + 0.553i$  as the ones that possibly coalesce first. We zoom into these eigenvalues in the top right picture. In the bottom left diagram, successive steps in the bisection method gives better approximation of the saddle point. Finally in the bottom right picture, we see that the saddle point was calculated at an accuracy at which the level sets of  $\underline{\sigma}_A$  are hard to compute.

There are other cases where the heuristic method fails to find the correct pair of eigenvalues whose components first coalesce.

$i$	$f(x_i)$	$M_i$	$\frac{M_i - f(x_i)}{f(x_i)}$	$ x_i - y_i $
1	<b>6.1325135002707E-4</b>	<b>6.1511092864335E-4</b>	3.03E-03	5.23E-03
2	<b>6.1511091521293E-4</b>	<b>6.1511092861426E-4</b>	2.18E-08	1.40E-05
3	<b>6.1511092861422E-4</b>	<b>6.1511092861423E-4</b>	3.35E-15	9.97E-10

TABLE 1. Convergence data for Example 8.1. Significant digits are in bold.

**Example 8.2.** Consider the matrix  $A$  generated by the following Matlab code:

```
A=zeros(10);
```

```
A(1:9,2:10)= diag([0.5330 + 0.5330i, 0.9370 + 0.1190i,...
    0.7410 + 0.8340i, 0.7480 + 0.8870i, 0.6880 + 0.6700i,...
    0.2510 + 0.7430i, 0.9540 + 0.6590i, 0.2680 + 0.6610i,...
    0.2670 + 0.4340i]);
```

```
A=      A+diag([0.9850 + 0.7550i,0.8030 + 0.7810i,...
    0.2590 + 0.5110i,0.3840 + 0.5310i,0.0080 + 0.5360i,...
    0.9780 + 0.2720i,0.7190 + 0.3100i,0.5560 + 0.8370i,...
    0.6350 + 0.7630i,0.5110 + 0.8870i]);
```

A sample run for this matrix is shown in Figure 8.2. The heuristic on minimal values of  $\underline{\sigma}_A$  on the edges of the Voronoi diagram identifies the top left and central eigenvalues as a pair for which the pseudospectral components first coalesce. However, the correct pair should be the central and bottom right eigenvalues.

Here are a few more observations. In our trials, we attempt to find the Wilkinson distance for bidiagonal matrices of size  $10 \times 10$  similar to the matrices in Examples 8.1 and 8.2. In all the examples we have tried, there was no need to perform step 1(a) of Algorithm 3.1 to achieve convergence to a critical point. The convergence for the matrix in Example 8.1 reflects the general performance of the (local) algorithm. As we have seen in Example 8.2, the heuristic for choosing a pair of eigenvalues may fail to choose the correct pseudospectral components which first coalesce as  $\epsilon$  increases. In a sample of 225 runs, we need to check other pairs of eigenvalues 7 times. In such cases, a different choice of a pair of eigenvalues still gave convergence to the Wilkinson distance, though whether this must always be the case is uncertain. The upper bounds for the critical value are also better approximates of the critical values than the lower bounds.

## 9. NON-LIPSCHITZ CONVERGENCE AND OPTIMALITY CONDITIONS

In this section, we discuss the convergence of Algorithm 2.1 in the non-Lipschitz case and give an optimality condition in step 2 of Algorithm 2.1. As one might expect in the smooth case in a Hilbert space, if  $x_i$  and  $y_i$  are closest points in the different components,  $\nabla f(x_i) \neq \mathbf{0}$  and  $\nabla f(y_i) \neq \mathbf{0}$ , then we have

$$\begin{aligned} x_i - y_i &= \lambda_1 \nabla f(y_i), \\ y_i - x_i &= \lambda_2 \nabla f(x_i). \end{aligned}$$



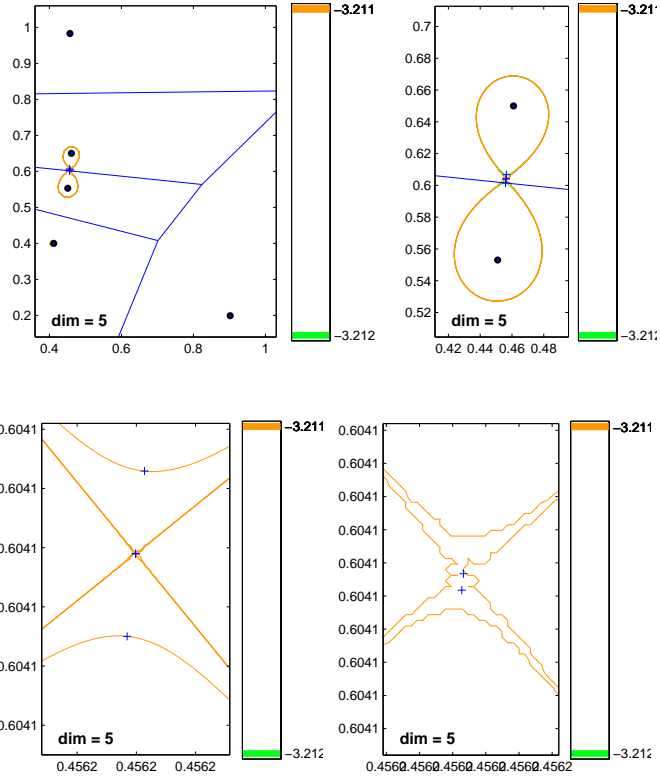


FIGURE 8.1. A sample run of Algorithm 3.1.

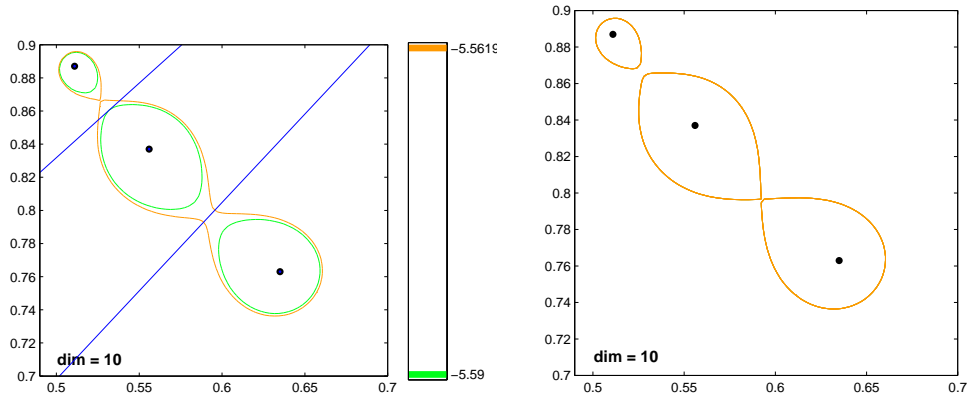


FIGURE 8.2. An example where the Voronoi diagram heuristic fails.

for  $\lambda_1, \lambda_2 > 0$ . The rest of this section extends this result to the nonsmooth case, making use of the language of variational analysis in the style of [36, 8, 14, 32] to describe the relation between subdifferentials of  $f$  and the normal cones of the level sets of  $f$ .

We now recall the definition of the Fréchet subdifferential, which is a generalization of the derivative to nonsmooth cases, and the Fréchet normal cone. A function  $f : X \rightarrow \mathbb{R}$  is *lsc* (lower semicontinuous) if  $\liminf_{x \rightarrow \bar{x}} f(x) \geq f(\bar{x})$  for all  $\bar{x} \in X$ .

**Definition 9.1.** Let  $f : X \rightarrow \mathbb{R} \cup \{+\infty\}$  be a proper lsc function. We say that  $f$  is *Fréchet subdifferentiable* and  $x^*$  is a *Fréchet-subderivative* of  $f$  at  $x$  if  $x \in \text{dom} f$  and

$$\liminf_{|h| \rightarrow 0} \frac{f(x+h) - f(x) - \langle x^*, h \rangle}{|h|} \geq 0.$$

We denote the set of all Fréchet-subderivatives of  $f$  at  $x$  by  $\partial_F f(x)$  and call this object the *Fréchet subdifferential* of  $f$  at  $x$ .

**Definition 9.2.** Let  $S$  be a closed subset of  $X$ . We define the *Fréchet normal cone* of  $S$  at  $x$  to be  $N_F(S; x) := \partial_F \iota_S(x)$ . Here,  $\iota_S : X \rightarrow \mathbb{R} \cup \{\infty\}$  is the indicator function defined by  $\iota_S(x) = 0$  if  $x \in S$ , and  $\infty$  otherwise.

Closely related to the Fréchet normal cone is the proximal normal cone.

**Definition 9.3.** Let  $X$  be a Hilbert space and let  $S \subset X$  be a closed set. If  $x \notin S$  and  $s \in S$  are such that  $s$  is a closest point to  $x$  in  $S$ , then any nonnegative multiple of  $x - s$  is a *proximal normal vector* to  $S$  at  $s$ . The set of all proximal normal vectors is denoted  $N_P(S; s)$ .

The proximal normal cone and the Fréchet normal cone satisfy the following relation. See for example [8, Exercise 5.3.5].

**Theorem 9.4.**  $N_P(S; \bar{x}) \subset N_F(S; \bar{x})$ .

Here is an easy consequence of the definitions.

**Proposition 9.5.** Let  $S_1$  be the component of  $\text{lev}_{\leq l_i} f$  containing  $x_0$  and  $S_2$  be the component of  $\text{lev}_{\leq l_i} f$  containing  $y_0$ . Suppose that  $x_i$  is a point in  $S_1$  closest to  $S_2$  and  $y_i$  is a point in  $S_2$  closest to  $x_i$ . Then we have

$$(y_i - x_i) \in N_P(\text{lev}_{\leq l_i} f; x_i) \subset N_F(\text{lev}_{\leq l_i} f; x_i).$$

Similarly,  $(x_i - y_i) \in N_F(\text{lev}_{\leq l_i} f; y_i)$ . These are two normals of  $\text{lev}_{\leq l_i} f$  pointing in opposite directions.

The above result gives a necessary condition for the optimality of step 2 in Algorithm 2.1. We now see how the Fréchet normals relate to the subdifferential of  $f$  at  $x_i$ ,  $y_i$  at  $\bar{z}$ . Here is the definition of the Clarke subdifferential for non-Lipschitz functions.

**Definition 9.6.** Let  $X$  be a Hilbert space and let  $f : X \rightarrow \mathbb{R}$  be a lsc function. Then the *Clarke subdifferential* of  $f$  at  $\bar{x}$  is

$$\partial_C f(\bar{x}) := \text{cl conv} \left\{ w - \lim_{i \rightarrow \infty} x_i^* \mid x_i^* \in \partial_F f(x_i), (x_i, f(x_i)) \rightarrow (\bar{x}, f(\bar{x})) \right\} + \partial_C^\infty f(\bar{x}),$$

where the *singular subdifferential* of  $f$  at  $\bar{x}$  is a cone defined by

$$\partial_C^\infty f(\bar{x}) := \text{cl conv} \left\{ w - \lim_{i \rightarrow \infty} \lambda_i x_i^* \mid x_i^* \in \partial_F f(x_i), (x_i, f(x_i)) \rightarrow (\bar{x}, f(\bar{x})), \lambda_i \rightarrow 0_+ \right\}.$$

For finite dimensional spaces, the weak topology is equivalent to the norm topology, so we may replace  $w - \lim$  by  $\lim$  in that setting. We will use the limiting subdifferential and the limiting normal cone, whose definitions we recall below, in the proof of the finite dimensional case of Theorem 9.11.

**Definition 9.7.** Let  $X$  be a Hilbert space and let  $f : X \rightarrow \mathbb{R}$  be a lsc function. Define the *limiting subdifferential* of  $f$  at  $\bar{x}$  by

$$\partial_L f(\bar{x}) := \{w - \lim_{i \rightarrow \infty} x_i^* \mid x_i^* \in \partial_F f(x_i), (x_i, f(x_i)) \rightarrow (\bar{x}, f(\bar{x}))\},$$

and the *singular subdifferential* of  $f$  at  $\bar{x}$ , which is a cone, by

$$\partial^\infty f(\bar{x}) := \{w - \lim_{i \rightarrow \infty} t_i x_i^* \mid x_i^* \in \partial_F f(x_i), (x_i, f(x_i)) \rightarrow (\bar{x}, f(\bar{x})), t_i \rightarrow 0_+\}.$$

The limiting normal cone is defined in a similar manner.

**Definition 9.8.** Let  $X$  be a Hilbert space and let  $S$  be a closed subset of  $X$ . Define the limiting normal cone of  $S$  at  $x$  by

$$N_L(S; x) := \{w - \lim_{i \rightarrow \infty} x_i^* \mid x_i^* \in N_F(S; x_i), S \ni x_i \rightarrow x\}.$$

It is clear from the definitions that the Fréchet subdifferential is contained in the limiting subdifferential, which is in turn contained in the Clarke subdifferential. Similarly, the Fréchet normal cone is contained in the limiting normal cone. We first state a theorem relating normal cones to subdifferentials in the finite dimensional case.

**Theorem 9.9.** [36, Proposition 10.3] *For a lsc function  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ , let  $\bar{x}$  be a point with  $f(\bar{x}) = \alpha$ . Then*

$$N_F(\text{lev}_{\leq \alpha} f; \bar{x}) \supset \mathbb{R}_+ \partial_F f(\bar{x}) \cup \{\mathbf{0}\}.$$

*If  $\partial_L f(\bar{x}) \not\ni \mathbf{0}$ , then also*

$$N_L(\text{lev}_{\leq \alpha} f; \bar{x}) \subset \mathbb{R}_+ \partial_L f(\bar{x}) \cup \partial^\infty f(\bar{x}).$$

The corresponding result for the infinite dimensional case is presented below.

**Theorem 9.10.** [8, Theorem 3.3.4] *Let  $X$  be a Hilbert space and let  $f : X \rightarrow \mathbb{R} \cup \{+\infty\}$  be a lsc function. Suppose that  $\liminf_{x \rightarrow \bar{x}} d(\partial_F f(x); \mathbf{0}) > 0$  and  $\xi \in N_F(\text{lev}_{\leq f(\bar{x})} f; \bar{x})$ . Then, for any  $\epsilon > 0$ , there exist  $\lambda > 0$ ,  $(x, f(x)) \in \mathbb{B}_\epsilon((\bar{x}, f(\bar{x})))$  and  $x^* \in \partial_F f(x)$  such that*

$$|\lambda x^* - \xi| \leq \epsilon.$$

With these preliminaries, we now prove our theorem for the convergence of Algorithm 2.1 to a Clarke critical point.

**Theorem 9.11.** *Suppose that  $f : X \rightarrow \mathbb{R}$ , where  $X$  is a Hilbert space and  $f$  is lsc. If  $\bar{z}$  is such that*

- (1)  $(\bar{z}, \bar{z})$  is a limit point of  $\{(x_i, y_i)\}_{i=1}^\infty$  in Algorithm 2.1, and
- (2)  $f$  is continuous at  $\bar{z}$ .

*Then one of these must hold:*

- (a)  $\bar{z}$  is a Clarke critical point,
- (b)  $\partial_C^\infty f(\bar{z})$  contains a line through the origin, or
- (c)  $\left\{ \frac{y_i - x_i}{|y_i - x_i|} \right\}_i$  converges weakly to zero.

*Proof.* We present both the finite dimensional and infinite dimensional versions of the proof to our result.

Suppose the subsequence  $\{(x_i, y_i)\}_{i \in J}$  is such that  $\lim_{i \rightarrow \infty, i \in J} (x_i, y_i) = (\bar{z}, \bar{z})$ , where  $J \subset \mathbb{N}$ . We can choose  $J$  so that none of the elements in  $\{(x_i, y_i)\}_{i \in J}$  are such that  $\liminf_{x \rightarrow x_i} d(\partial_F f(x); \mathbf{0}) = 0$  or  $\liminf_{y \rightarrow y_i} d(\partial_F f(y); \mathbf{0}) = 0$ , otherwise we have  $\mathbf{0} \in \partial_C f(\bar{z})$  by the definition of the Clarke subdifferential, which is what we seek to prove. (In finite dimensions, the condition  $\liminf_{x \rightarrow x_i} d(\partial_F f(x); \mathbf{0}) = 0$  can be replaced by  $\mathbf{0} \in \partial_L f(x_i)$ .) We proceed to apply Theorem 9.10 (and Theorem 9.9 for finite dimensions) to find out more about  $N_F(\text{lev}_{\leq l_i} f; x_i)$ .

We first prove the result for finite dimensions. If  $\mathbf{0} \in \partial_L f(\bar{z})$ , we are done. Otherwise, by Proposition 9.5 and Theorem 9.9, there is a positive multiple of  $v = \lim_{i \rightarrow \infty} \frac{y_i - x_i}{|y_i - x_i|}$  that lies in either  $\partial_L f(\bar{z})$  or  $\partial^\infty f(\bar{z})$ . Similarly, there is a positive multiple of  $-v = \lim_{i \rightarrow \infty} \frac{x_i - y_i}{|y_i - x_i|}$  lying in either  $\partial_L f(\bar{z})$  or  $\partial^\infty f(\bar{z})$ . If either  $v$  or  $-v$  lies in  $\partial_L f(\bar{z})$ , then we can conclude  $\mathbf{0} \in \partial_C f(\bar{z})$  from the definitions. Otherwise both  $v$  and  $-v$  lie in  $\partial_C^\infty f(\bar{z})$ , so  $\mathbb{R}\{v\} \subset \partial_C^\infty f(\bar{z})$  as needed.

We now prove the result for infinite dimensions. The point  $\bar{z}$  is the common limit of  $\{x_i\}_{i \in J}$  and  $\{y_i\}_{i \in J}$ . By the optimality of  $|x_i - y_i|$  and Proposition 9.5, we have  $y_i - x_i \in N_F(\text{lev}_{\leq l_i} f; x_i)$  and  $x_i - y_i \in N_F(\text{lev}_{\leq l_i} f; y_i)$ . By Theorem 9.10, for any  $\kappa_i \rightarrow 0_+$ , there is a  $\lambda_i > 0$ ,  $x'_i \in \mathbb{B}_{\kappa_i |x_i - y_i|}(x_i)$  and  $x_i^* \in \partial_F f(x'_i)$  such that  $|\lambda_i x_i^* - (y_i - x_i)| < \kappa_i |y_i - x_i|$ . Similarly, there is a  $\gamma_i > 0$ ,  $y'_i \in \mathbb{B}_{\kappa_i |y_i - x_i|}(y_i)$  and  $y_i^* \in \partial_F f(y'_i)$  such that  $|\gamma_i y_i^* - (x_i - y_i)| < \kappa_i |x_i - y_i|$ . If either  $x_i^*$  or  $y_i^*$  converges to  $\mathbf{0}$ , then  $\mathbf{0} \in \partial_C f(\bar{z})$ , and we are done. Otherwise, by the Banach Alaoglu theorem, the unit ball is compact, so  $\left\{ \frac{1}{|x_i^*|} x_i^* \right\}_i$  and  $\left\{ \frac{1}{|y_i - x_i|} (y_i - x_i) \right\}_i$  have weak cluster points. We now show that they must have the same cluster points by showing that their difference converges to  $\mathbf{0}$  (in the strong topology). Now,

$$\begin{aligned} \left| \frac{\lambda_i x_i^*}{|y_i - x_i|} \right| &\leq \left| \frac{\lambda_i x_i^*}{|y_i - x_i|} - \frac{y_i - x_i}{|y_i - x_i|} \right| + \left| \frac{y_i - x_i}{|y_i - x_i|} \right| \\ &\leq \kappa_i + 1, \end{aligned}$$

and similarly,  $1 - \kappa_i \leq \left| \frac{\lambda_i x_i^*}{|y_i - x_i|} \right|$ , so  $\left| \frac{\lambda_i x_i^*}{|y_i - x_i|} \right| \rightarrow 1$ , and thus

$$\left| \frac{\lambda_i x_i^*}{|y_i - x_i|} - \frac{x_i^*}{|x_i^*|} \right| = \left\| \frac{\lambda_i x_i^*}{|y_i - x_i|} - \frac{x_i^*}{|x_i^*|} \right\| \rightarrow 0.$$

This means that

$$\left| \frac{x_i^*}{|x_i^*|} - \frac{y_i - x_i}{|y_i - x_i|} \right| \leq \left| \frac{\lambda_i x_i^*}{|y_i - x_i|} - \frac{x_i^*}{|x_i^*|} \right| + \left| \frac{\lambda_i x_i^*}{|y_i - x_i|} - \frac{y_i - x_i}{|y_i - x_i|} \right| \rightarrow 0,$$

which was what we claimed earlier. This implies that  $\frac{x_i^*}{|x_i^*|}$  and  $\frac{y_i^*}{|y_i^*|}$  have weak cluster points that are the negative of each other.

We now suppose that conclusion (c) does not hold. If  $\{x_i^*\}_i$  has a nonzero weak cluster point, say  $\bar{x}^*$ , then  $\bar{x}^*$  belongs to  $\partial_C f(\bar{z})$ . Then  $\{y_i^*\}_i$  either has a weak cluster point  $\bar{y}^*$  that is strictly a negative multiple of  $\bar{x}^*$ , which implies that  $\mathbf{0} \in \partial_C f(\bar{z})$  as claimed, or there is some  $\bar{y}^{*,\infty} \in \partial_C^\infty f(\bar{z})$  which is a negative multiple of  $\bar{x}^*$ , which also implies that  $\mathbf{0} \in \partial_C f(\bar{z})$  as needed.

If neither  $\{x_i^*\}_i$  or  $\{y_i^*\}_i$  converges weakly, then two (nonzero) weak cluster points of  $\frac{x_i^*}{\|x_i^*\|}$  and  $\frac{y_i^*}{\|y_i^*\|}$  that point in opposite directions give a line through the origin in  $\partial_C^\infty f(\bar{z})$  as needed.  $\square$

In finite dimensions, conclusion (b) of Theorem 9.11 is precisely the lack of “epi-Lipschitzness” [36, Exercise 9.42(b)] of  $f$ . One example where Algorithm 2.1 does not converge to a Clarke critical point but to a point with its singular subdifferential  $\partial_C^\infty f(\cdot)$  containing a line through the origin is  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $f(x) = -\sqrt{|x|}$ . Algorithm 2.1 converges to the point 0, where  $\partial_C f(0) = \emptyset$  and  $\partial_C^\infty f(0) = \mathbb{R}$ . We do not know of an example where only condition (c) holds.

# ACKNOWLEDGMENTS

We thank Jianxin Zhou for comments on an earlier version of the manuscript, and we thank an anonymous referee for feedback, which have improved the presentation in the paper.

# REFERENCES

- [1] R. Alam and S. Bora, *On sensitivity of eigenvalues and eigendecompositions of matrices*, Linear Algebra Appl., 396 (2005), pp. 273-301.
- [2] R. Alam, S. Bora, R. Byers and M.L. Overton, *Characterization and construction of the nearest defective matrix via coalescence of pseudospectral components*, submitted, 2009.
- [3] A. Ambrosetti, *Critical points and nonlinear variational problems*, Mémoires de la Société Mathématique de France, Sér. 2, 49 (1992), p. 1-139.
- [4] A. Ambrosetti and P. Rabinowitz, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349-381.
- [5] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley 1984. Reprinted by Dover 2007.
- [6] V. Barutello and S. Terracini, *A bisection algorithm for the numerical mountain pass*, Non-linear differ. equ. appl. 14 (2007) 527-539.
- [7] R. Benedetti & J.-J. Risler, *Real algebraic and semi-algebraic sets* (Hermann, Paris, 1990).
- [8] J. M. Borwein and Q. J. Zhu, *Techniques of Variational Analysis*, Springer, 2005.
- [9] S. Boyd and V. Balakrishnan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its  $L_\infty$ -norm*, Systems and Control Letters 15 (1990) 1-7.
- [10] J.V. Burke, A.S. Lewis and M.L. Overton. *Spectral conditioning and pseudospectral growth*. Numerische Mathematik, 107:27-37, 2007
- [11] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 875-881.
- [12] Kung-Ching Chang, *Variational methods for non-differentiable functionals and their applications to partial differential equations*, Journal of Mathematical Analysis and its Applications, 80, 102-129 (1981).
- [13] Y.S. Choi and P. J. McKenna, *A mountain pass method for the numerical solution of semi-linear elliptic problems*, Nonlinear Anal., 20 (1993), pp. 417-437.
- [14] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983. Republished as Vol. 5, Classics in Applied Mathematics, SIAM, 1990.
- [15] M. Coste, *An Introduction to O-minimal Geometry*, Instituti Editoriali e poligrafici internazionali (Universita di Pisa, 1999), available electronically at <http://perso.univ-rennes1.fr/michel.coste/>
- [16] M. Coste, *An Introduction to Semialgebraic Geometry*, Instituti Editoriali e poligrafici internazionali (Universita di Pisa, 2002), available electronically at <http://perso.univ-rennes1.fr/michel.coste/>
- [17] M. Degiovanni and M. Marzocchi, *A critical point theory for nonsmooth functionals*, Ann. Math. Pura. Appl. 167 (1994), pp. 73-100

- [18] J. W. Demmel, *On condition numbers and the distance to the nearest ill-conditioned problem*, Numerische Mathematik, 51, 251-289, 1987.
- [19] Zhonghai Ding, David Costa and Goong Chen, *A high-linking algorithm for sign-changing solutions of semilinear elliptic equations*, Nonlinear Analysis 38 (1999) 151-172.
- [20] L. van den Dries, *Tame Topology and o-minimal Structures* (Cambridge, 1998).
- [21] G Henkelman, G Jóhannesson, H Jónsson, *Methods for finding saddle points and minimum energy paths*, In: Progress in Theoretical Chemistry and Physics. S.D. Schwartz (ed.) Vol. 5, Kluwer 2000.
- [22] J. Horák, *Constrained mountain pass algorithm for the numerical solution of semilinear elliptic problems*, Numerische Mathematik 98 (2004) 251-276.
- [23] A.D. Ioffe and E. Scwhartzman, *Metric critical point theory 1: Morse regularity and homotopic stability of a minimum*, J. Math Pures Appl. 75 (1996), pp. 125-153.
- [24] Youssef Jabri, *The Mountain Pass Theorem*, Cambridge, 2003.
- [25] G. Katriel, *Mountain pass theorem and a global homeomorphism theorem*, Ann. Institut Henri Poincaré, Analyse Non Linéaire, 11 (1994), pp. 189-209.
- [26] Yongxin Li and Jianxin Zhou, *A minimax method for finding multiple critical points and its applications to semilinear PDES*, SIAM J. Sci. Comput., Vol 23, No. 3 , pp 840-865, 2001.
- [27] Yongxin Li and Jianxin Zhou, *Convergence results of a local minimax method for finding multiple critical points*, SIAM J. Sci. Comput., Vol 24, No. 3, pp. 865-885, 2002.
- [28] A. N. Malyshev, *A formula for the 2-norm distance from a matrix to the set of matrices with multiple eigenvalues*, Numer. Math. 83 (1999) 443-454.
- [29] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer, Berlin, 1989.
- [30] E. Mengi, 2009. private communication.
- [31] E. Mengi and M. Overton, *Algorithms for the computation of the pseudospectral radius and the numerical radius of a matrix*, IMA J. Numer. Anal. (2005) 25, 648-669.
- [32] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation I and II*, Springer, Berlin, 2006.
- [33] J. J. Moré and T. S. Munson, *Computing mountain passes and transition states*, Math. Program. Ser. B 100: 151-182 (2004).
- [34] L. Nirenberg, *Variational Methods in Nonlinear Problems. Topics in the calculus of variations (Montecatini Terme, 1987)*, 100-119, Lectures Notes in Mathematics, 1365, Springer, 1989.
- [35] P.H. Rabinowitz, *Minimax Methods in Critical Point Theory with Applications to Differential Equations*, CBMS Regional Conference ser. Math, AMS, 65, 1986.
- [36] R.T. Rockafellar and R. J-B Wets, *Variational Analysis*, Springer, 1998.
- [37] S. Shi, *Ekeland's variational principle and the mountain pass lemma*, Acta. Math. Sin., (N.S.), 1, no. 4, 348-355 (1985).
- [38] J.E. Sinclair and R. Fletcher, *A new method of saddle-point location for the calculation of defect migration energies*, J. Phys. C: Solid State Phys., pp 864-870, Vol 7, 1974.
- [39] M. Struwe, *Variational Methods* (3<sup>rd</sup> edition) (Springer, 2000).
- [40] L.N. Trefethen and M. Embree, *Spectra and Pseudospectra*, Princeton, NJ, 2005.
- [41] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford, 1965.
- [42] M. Willem, *Un Lemme de déformation quantitatif en calcul des variations*. (French) [*A quantitative deformation lemma in the calculus of variations*.] Institut de Mathématiques pures et appliquées [Applied and Pure Mathematics Institute], Recherche de mathématiques [Mathematics Research] no. 19, Catholic University of Louvain, May 1992.
- [43] T. G. Wright, EigTool: a graphical tool for nonsymmetric eigenproblems, 2002; available online at <http://web.comlab.ox.ac.uk/pseudospectra/eigtool/>
- [44] Xudong Yao and Jianxin Zhou, *A local minimax characterization of computing multiple nonsmooth saddle critical points*, Math. Program., Ser. B 104, 749-760 (2005).
- [45] Xudong Yao and Jianxin Zhou, *Unified convergence results on a minimax algorithm for finding multiple critical points in Banach spaces*, SIAM J. Num. Anal., 45 (2007) 1330-1347.

*Current address:* School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853.

*E-mail address:* `aslewis@orie.cornell.edu`.

*Current address:* Combinatorics and Optimization, University of Waterloo, 200 University Ave W., Waterloo, ON, Canada N2l 3G1.

*E-mail address:* `chj2pang@math.uwaterloo.ca`